REPORT DOCUMENTATION PAGE

ved
-0188

AFRL-SR-BL-TR-99-
0322

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
8/31/98

3. REPORT TYPE AND DATES COVERED
Progress Report for 8/97 to 8/98

4. TITLE AND SUBTITLE
Wireless Connectivity to ATM Communication Grid

5. FUNDING NUMBERS
F49620-97-1-0471

6. AUTHOR(S)
Dr. Veeramuthu Rajaravivarma
Dr. Krishna Sivalingam

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
North Carolina A&T State University
1601 E. Market Street
Greensboro, NC 27411

8. PERFORMING ORGANIZATION REPORT NUMBER
4-41145

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
Attn: Dr. Jon Sjogren, Program Manager
Dept. of the Air Force AFSOR 64-1
110 Duncan Avenue, Room B115
Bolling Air Force Base, DC 20332-8080

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A network research laboratory has been established by Dr. Sivalingam in the School of Electrical Engg. & Computer Science at Washington State University, Pullman. The AFOSR funds were used to purchase a 12-port Fore ATM switch, ATM network interface cards, a SUN UltraSPARC workstation, Lucent WavePoint wireless bridge, and Lucent WaveLAN wireless network interface cards. At North Carolina A&T State University, a $8 million new technology building is under construction. AFSOR funds were used to purchase Fore ATM switches, ATM network interface cards, Fiber patch card, AMP wireless access point, AMP wireless card, Gateway PC/TV Destination, HP Scanner, and two Data cabinets. All these equipments will be used by undergraduate students.

We have designed and analyzed a low-power access protocol, called ECMAC (Energy conserving medium access control). This protocol has been simulated using a freely available discrete-event simulation package. Performance results from the analysis show that this protocol does have better energy consumption characteristics compared to a number of other protocols, including the IEEE 802.11 standard. The next stage in this research is to implement the MAC protocol in software and reconfigurable hardware for real-time power analysis. We have made significant modifications to the RSVP protocol, the Internet standard for reservation signaling.

14. SUBJECT TERMS

15. NUMBER OF PAGES
153

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

Under Review with ACM/Baltzer Mobile Networking and Applications Journal, Submitted March 1998.

ACCEPTED, ~~MARCH~~ 1998, JUNE.

Design and Analysis of Low-Power Access Protocols for Wireless and Mobile ATM Networks

Krishna M. Sivalingam[1], Jyh-Cheng Chen[2], Prathima Agrawal[3] and Mani B. Srivastava[4] [*]

[1]School of Electrical Engineering & Computer Science, Washington State University, Pullman, WA 99164

[2]Department of Electrical & Computer Engineering, State University of New York at Buffalo, Buffalo, NY 14260

[3]Networked Computing Technology Department, AT&T Labs, Whippany, NJ 07981

[4]Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA 90095

Abstract

This paper describes the design and analysis of a low-power medium access control (MAC) protocol for wireless/mobile ATM networks. The protocol – denoted EC-MAC (energy conserving medium access control) – is designed to support different traffic types with quality-of-service (QoS) provisions. The network is based on the infrastructure model where a base station (BS) serves all the mobiles currently in its cell. A reservation-based approach is proposed, with appropriate scheduling of the requests from the mobiles. This strategy is utilized to accomplish the dual goals of reduced energy consumption and quality of service provision over wireless links. A *priority round robin with dynamic reservation update and error compensation* scheduling algorithm is used to schedule the transmission requests of the mobiles. Discrete-event simulation has been used to study the performance of the protocol. A comparison of energy consumption of the EC-MAC to a number of other protocols is provided. This comparison indicates the EC-MAC has in general better energy consumption characteristics. Performance analysis of the proposed protocol with respect to different quality-of-service parameters using video, audio and data traffic models is provided.

1 Introduction

Wireless services, such as cellular voice, PCS (Personal Communication Services), mobile data, and wireless LANs, are anticipated as some of the strongest growth areas in telecommunications [1,2]. Third-generation networks designed to carry multimedia traffic such as voice, video, audio, animation, images, and data transmission are under intensive research investigation. The goal of the wireless networking research is to provide seamless communications, high bandwidth, and guaranteed quality-of-service regardless of location and mobility constraints.

[*] CONTACT: Krishna Sivalingam, School of Electrical Engg. and Computer Science, Washington State University, Pullman, WA 99164. Phone: (509) 335-3220, Fax: (509) 335-3818. Email: `krishna@eecs.wsu.edu`. The authors can be reached at krishna@eecs.wsu.edu, jcchen@eng.buffalo.edu, pa@research.att.com, and mbs@janet.ucla.edu. Part of the research was done while the author was with AT&T Labs and the University of North Carolina at Greensboro. Part of the research was supported by Air Force Office of Scientific Research grant F-49620-97-1-0471.

The bandwidth offered by the wireless network will typically tend to lag behind that offered by the wired network. The wired network will serve as the primary or backbone system with enormous bandwidth, while the wireless network will extend the reach of the network. In order to avoid a serious mismatch between future wired and wireless networks, broadband wireless systems should offer similar services as the current and proposed wired broadband networks. These wired broadband systems, such as B-ISDN ATM [3], are expected to offer constant bit-rate (CBR), variable bit-rate (VBR), and available bit-rate (ABR) services designed to support multimedia applications [4].

Research efforts to study the integration of wireless and ATM networks have been reported in [5–7]. The objective of integrating ATM and wireless networks is to ensure that the services offered by the wired ATM network are seamlessly extended to the mobile and wireless users. A number of interesting challenges arise as a result of this integration that are addressed in [5–7]. The importance of this topic has led to the establishment of a Wireless ATM working group within the ATM Forum [8].

One of the fundamental challenges in extending the ATM network to the wireless domain is to extend the virtual circuit (VC) service with quality-of-service to mobile connections. We believe that some of the support for this functionality needs to be provided at the wireless media access layer. Traditional access protocols for wireless networks do not consider quality-of-service issues or diverse traffic types as envisioned for multimedia networks. To provide CBR, VBR, and ABR services to end users, a wireless medium access control (MAC) protocol must be able to provide bandwidth on demand with different levels of service. The high error rates of wireless networks may preclude definite guarantees of service. The network could offer different levels of service and mobile applications may adapt to the offered service quality as required. This paper also considers the important dimension of reduced energy consumption due to MAC-related activities at the mobiles. Mobile battery power is limited and therefore power consumption should be minimal.

The objective of this paper is to present the design and analysis of a media access protocol, referred to as EC-MAC (Energy-Conserving MAC). The protocol design is driven by two major factors. The first factor is that the access protocol should be energy-efficient since the mobiles typically have limited power capacity. The second factor is that the protocol should provide support for multiple traffic types, with appropriate quality-of-service levels for each type. In related work, quality-of-service issues in wireless networks have been considered separately in broadband wireless networks [9, 10]. Reduced energy consumption at the MAC layer has been considered in [11]. Our goal is to define a comprehensive access protocol combining these factors that will form the basis for future broadband wireless networks.

The goals of low energy consumption and QoS provision lead us to a protocol that is based on reservation and scheduling strategies. Sharing of the wireless channel among multiple mobiles and connections requires that some form of statistical multiplexing be used. The *base station* (BS) receives transmission requests or VC-setup requests from the mobiles. The base station schedules the time slots on the channels to the mobiles based on this information. The key to providing service quality will be the scheduling algorithm executed at the base station. Previous work has reported mechanisms for providing the base station with the transmission requests [12, 13]. Scheduling algorithms for wireless networks are described in [14, 15]. The new features of the protocol design described in this paper are that it considers low-power operation, multiple traffic types, error state of mobiles, and provides service quality with respect to offered bandwidth.

The proposed protocol is evaluated by discrete-event simulation where voice traffic is modeled by a slow speech activity detector (SAD) for talkspurts and silent gaps [16]. Video traffic is modeled as a H.263 source [17, 18] obtained from traffic traces. Data is modeled as self-similar traffic obtained from [19, 20]. Various QoS parameters for voice, video, and data traffic with varying number of mobiles in a cell are considered. A comparison of energy consumption for EC-MAC and other protocols shows that the energy consumption of EC-MAC is independent of the traffic load since collisions are minimized. As the traffic load increases, EC-MAC consumes less energy than other protocols. The results also indicate that EC-MAC achieves high channel utilization.

The rest of the paper is organized as follows. Section 2 describes the network architecture. Section 3 briefly describes some of the low power access protocol design issues in wireless networks. Section 4 describes the traffic characteristics, the access protocol and the scheduling algorithm. Section 5 provides the performance analysis of the protocol with voice, video, and data traffic. Section 6 summarizes the paper.

2 Network Architecture

The network architecture is derived from the SWAN network built at Bell Labs [21] – one of the first wireless ATM network testbeds. The wired backbone network is comprised of a hierarchy of wide-area and local-area ATM networks, with wireless links being used to provide last hop access. In addition to connecting conventional wired server hosts and client end-points, the wired backbone also connects to special switching nodes called *base stations*. The geographical area for which a base station radio port acts as the gateway is called its *cell*[†].

Figure 1 shows the functional blocks in the wireless last hop. The primary function of the base station is to switch cells among various wired and wireless ATM adapters attached to the base station under the control of a Connection Manager signaling module. The base station is effectively an ATM switch that has wireless (RF) ATM adapters on some of its ports. At the other end of the wireless last hop is the mobile that has a RF wireless adapter, a connection signaling manager module, and a module that routes cells from/to various software/hardware agents acting as sinks and sources of ATM cells within the mobile. The connection managers at the mobile and the base stations implement, among other things, the VC re-routing protocols required to handle mobility.

The subset of the wireless last hop that is of relevance in making ATM wireless is the shaded area in the picture – streams of ATM cells belonging to different VCs, with different QoS requirements, from the higher

[†] Please note the term *cell* here is different from the term cell used to denote the basic transmission unit in ATM. The difference will be apparent based on the context.

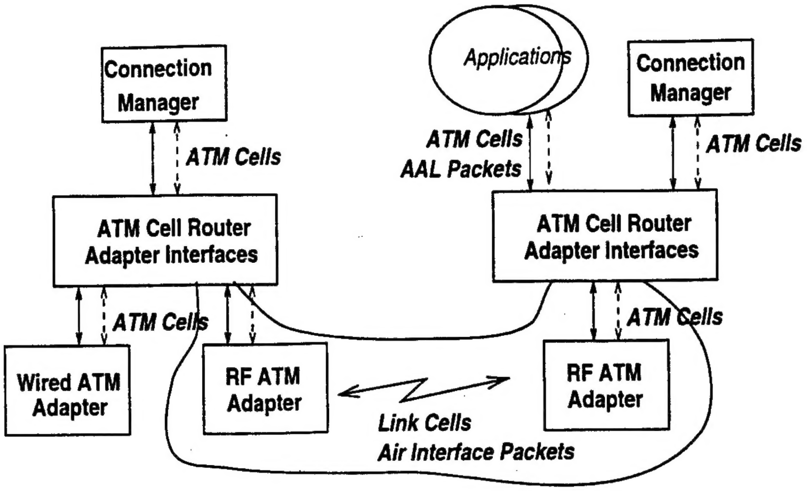


Figure 1: The Wireless Last Hop.

level ATM layers need to be multiplexed across the wireless link between the mobile and its base station. Whereas in the wired scenario each host has a dedicated point-to-point link to the corresponding switch port, the wireless case requires the air resources to be shared among the various mobiles located in a cell and communicating to the switch (that is, the base station). The need to take QoS into account necessitates a unification of the normally disjoint functions of cell multiplexing and medium access control. At the time of initial VC set-up, the connection manager in the ATM layer contacts the MAC subsystem to perform an admission control check based on the quality-of-service requirements of the connection. Similarly, following a hand-off, the connection manager at a base station contacts the MAC subsystem to verify if the quality-of-service requirements of the newly routed VC can be supported.

The following sections describe ATM and wireless integration issues, and the design of the protocol.

3 Low power access protocol design

The previous section described the network architecture. This section outlines the main factors that were influential in our protocol design.

Mobile computers typically have limited energy for computing and communications because of the short battery lifetimes. Conserving battery power in mobiles should be a crucial consideration in designing protocols for mobile computing. This issue should be considered through all layers of the protocol stack, including the application layer. Low-power design at the hardware layers uses different techniques including variable clock speed CPUs, flash memory, or disk spindowns. At the application layer, low-power video compression, transcoding at the base station and energy efficient database operation have been considered. In [22], the power drained by the network interface in hand-held devices was studied. An energy efficient probing scheme for error control in link layer is proposed in [23, 24]. The interaction of error control and forward error correction schemes in the link layer are studied from energy efficiency perspective in [25].

The chief sources of energy consumption in the mobile unit due to MAC related activities are the CPU, the transmitter, and the receiver. Mobile CPU usage may be reduced by relegating most of the high-complexity computation (related to media access) to the stationary network. Therefore, the focus of this work is on transceiver usage. The radio can operate in three modes: standby, receive, and transmit. In general, the radio consumes more power in the transmit mode than in the receive mode, and consumes least power in the standby mode. For example, the Proxim RangeLAN2 2.4 Ghz 1.6 Mbps PCMCIA card requires 1.5W in transmit, 0.75W in receive, and 0.01W in standby mode. In addition, turnaround between transmit and receive modes (and vice-versa) typically takes between 6 to 30 microseconds. Also, power consumption for Lucent's 15 dBm 2.4 GHz 2 Mbps Wavelan PCMCIA card is 1.82W in transmit mode, 1.80W in receive mode, and 0.18W in standby mode. Similar figures are 3.0W, 1.48W, and 0.18W, respectively for a 24.5 dBm 915 MHz 2 Mbps PCMCIA card. The power consumption will be higher for higher bit rates due to the higher equalization complexity.

The objective of MAC protocol design should be to minimize energy consumption while maximizing protocol performance. The following are some principles that may be observed to conserve energy at the MAC level.

1. Collision should be eliminated as far as possible since it results in retransmissions that leads to unnecessary energy consumption and also to possibly unbounded delays. Note that retransmissions cannot be completely avoided due to the high link error-rates and due to user mobility. For example, new users registering with the base station may have to use some form of random access protocol. However, using a small packet size for registration and bandwidth requests can reduce energy consumption.

   Techniques such as reservation and polling can help meet the requirement that collisions be minimized. Reservation and polling based protocols for wireless ATM networks have been proposed in [13,26] and [21,27], respectively.

2. In a typical wireless broadcast environment, the receiver has to be powered on at all times resulting in significant energy consumption. The receiver subsystem typically receives all packets and forwards only the packets destined for this mobile. For instance, this is the default mechanism used in IEEE 802.11 where the receiver is expected to keep track of channel status through constant monitoring.

   One possible way to reduce receiver power-on time is to broadcast a data transmission schedule for each mobile. This will enable a mobile to switch to standby mode until its alloted slots. This approach has been described in [11,26].

3. Significant time and energy is spent by the mobile radio in switching from transmit to receive modes, and vice-versa. This turnaround is a crucial factor in the performance of the protocol. A protocol such as DQRUMA [12] that allocates permission on a slot-by-slot basis will suffer significant overhead due to turnaround. This protocol allocates permission for the current slot and expects the mobile to

turn on the receiver at the start of the next slot to determine the new allocation. In order to reduce turnaround, a mobile should be allocated contiguous slots for transmission and reception whenever possible.

4. The IEEE 802.11 standard recommends the following technique for power conservation. A mobile that wishes to conserve power may switch to sleep mode and inform the base station of this decision. From that point on, the base station buffers packets destined for this mobile. The base station periodically transmits a beacon that contains information about such buffered packets. Upon waking up, the mobile listens for this beacon and informs the base station that it is ready to receive. The base station then forwards the buffered packets to the mobile station. This approach conserves power at the mobile but results in additional delays that may affect quality-of-service (QoS). It is essential to quantify this delay in the presence of QoS delay bounds for individual VCs.

5. The HIPERLAN standard for wireless LANs [28] provides two types of power saving mechanisms. The implicit mechanism turns on the equalizer only when the mobile is the intended destination of the downlink packet. The explicit mechanism allows the mobile to receive only during pre-arranged intervals instead of continuously. A mobile entering power-saver state informs a power-supporter mobile (possibly with a infrastructure power source) of the periodicity and length of the duration during which the mobile will be turned on. The p-supporter station receives and stores packets addressed for the power-saver mobiles it is supporting. This is essentially a distributed version of the 802.11 mechanism.

6. If reservations are used to request bandwidth, it will be more efficient (power-wise and bandwidth-wise) to request multiple cells with a single reservation packet. For example, an ATM-aware MAC layer can request resources for a complete or partial AAL5 packet instead of a cell-by-cell basis. This suggests that the mobile should request larger chunks of bandwidth to reduce the reservation overhead leading to better bandwidth and energy consumption efficiency. Some of the earlier protocols utilize such reservation modes for CBR traffic. For more dynamic VBR traffic, occasional queue status updates are utilized to inform the base station of changing traffic needs [26, 27].

7. Assume that mobiles transmit requests and that the base station uses a scheduling algorithm to allocate slots as in [12, 13, 26, 27]. A distributed algorithm where each mobile computes the schedule independently may not be desirable because: (i) it may not receive all the reservation requests due to radio and error constraints, and (ii) schedule computation consumes energy and is thus better relegated to the base station. This suggests that a centralized scheduling mechanism will be more energy efficient.

The issues listed above have influenced the design of the proposed protocol. A comparison of some of the earlier protocols from an energy consumption perspective is presented in detail in [29]. A summary of the results is presented in Section 5.1.

4 Protocol Description

This section describes the traffic types the protocol is designed for, the design decisions involved, and the access protocol.

The ATM Forum Traffic Management (TM) 4.0 specifications identifies five service categories: (i) CBR – Constant bit rate, (ii) rt-VBR – Real-time variable bit rate, (iii) nrt-VBR – Non-real-time variable bit rate, (iv) UBR – Unspecified bit rate and (v) ABR – Available bit rate.

As described in [4], CBR and UBR are the simpler categories. Voice and video could be handled by the network as CBR traffic allocating a fixed amount of bandwidth at regular intervals. However, video sources and voice with speech activity detection sources are better characterized as VBR. Typical computer communication messages can be modeled as UBR traffic. More complex categories are the ABR, rt-VBR, and nrt-VBR.

The proposed protocol considers three types of traffic: CBR, VBR, and UBR. The relevant source models that are used as examples of these traffic types are described in Section 5.2.

4.1 Protocol definition

The previous sections outlined the traffic types, the low power and QoS factors influencing protocol design. The proposed protocol is defined in this section.

A mobile can originate and terminate multiple connections that enable it to communicate with other computers and communication devices. All communication to and from the mobile is through the BS. Each such connection is referred to as a Virtual Circuit (VC). Each VC is associated with a transmission priority established by the mobile application utilizing this VC for communication. These priorities will be utilized by the BS when allocating channels to the mobiles. Each mobile maintains a separate queue for each of its VCs, as shown in Figure 2. Information arrives at each queue in the form of a *packet* and is buffered until transmission.

As described earlier, one of the objectives of extending ATM network services to wireless networks is to provide end-to-end service quality. In wired networks, QoS is achieved using appropriate scheduling algorithms at the ATM switches. A similar principle has to be applied to the wireless network where the base station is an extended ATM switch with wireless and mobility support. The general consensus observed in recent research on wireless ATM networks is that some form of reservation combined with a scheduling mechanism should be provided at the MAC level [30]. Consistent with this idea, the protocol we propose is based on using a scheduling algorithm to allocate bandwidth to the VCs.

The access protocol is defined here for an infra-structure network with a single base station serving mobiles in its coverage area. In order to extend this protocol to an ad-hoc network, the mobiles could elect a co-

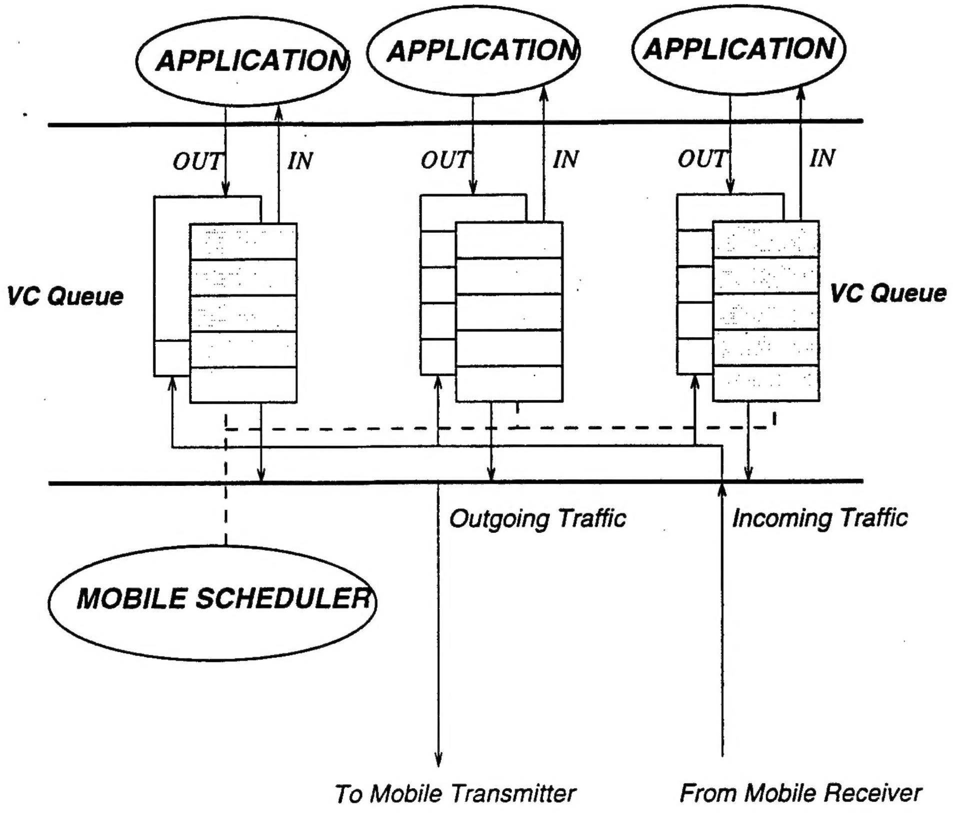


Figure 2: Virtual Circuit Queue structure within a mobile.

ordinator that will perform the functions of the base station. Issues such as coordinator moving away and downtime need to be carefully addressed, and are beyond the scope of this paper.

Each registered mobile is represented by an unique MACid that may be reassigned after handoff to a new base station. Every VC in every mobile is given a VCid – unique within the mobile. The <MACid,VCid> pair represents a unique VC within a mobile. This is similar to the combination of IP addresses and port numbers in TCP/IP networks.

Transmission in EC-MAC is organized by the base station into frames. Each frame is composed of a fixed number of slots, where each slot equals the basic unit of wireless data transmission. The basic unit of transmission in the wired ATM network is defined to be 53 bytes. In the wireless network, a different slot size may be used. For example, the SWAN testbed uses a slot-size of 64 bytes that arises due to current hardware limitations. The extra bytes may be used for wireless link header and control information.

The frame is divided into multiple phases as shown in Fig. 3:

Frame Synchronization: At the start of each frame, the BS transmits the frame synchronization message (FSM) on the downlink. This message contains framing and synchronization information, the uplink transmission order for reservations, and the number of slots in the new user phase.

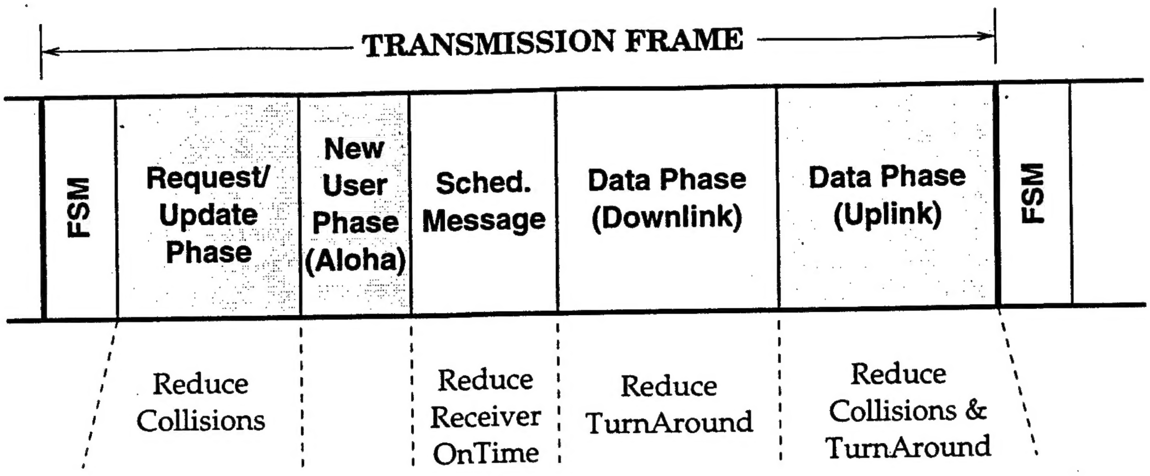


Figure 3: Definition of the different phases in EC-MAC protocol.

From battery power conservation perspective, it is desirable that the request/update phase (following this phase) should not operate in a contention mode. The request/update phase can be made collisionless by letting the base station broadcast a list containing the set of the mobile IDs. The transmission order of the IDs implicitly defines the order in which mobiles transmit their request/update information. Each mobile is allocated one slot during the request/update phase. Studies in the performance analysis section show that the maximum number of mobiles that can be supported using a single channel is of the order of a few tens of mobiles. Therefore, the overhead incurred in broadcasting this identifier list during the frame synchronization is of the order of a few hundreds of bytes.

Request/Update: The request/update phase is composed of uplink request transmissions from the mobiles. During the uplink phase, each registered mobile transmits new connection requests and queue status of established VBR and UBR queues. In case of CBR traffic, a specified number of slots are reserved in every frame for a specific VC until the mobile indicates that it is done with the VC transmission.

New-User Phase: This phase allows new mobiles to register with the base station. This phase is operated in a contention mode, using Slotted Aloha. The length of this phase is variable. The base station FSM broadcasts the available number of slots for user registration during this phase. The base station initially starts with a small number of slots, and dynamically adjusts the number of slots based on monitoring the number of collisions. A maximum number of new-user slots is specified. The base station transmits all the acknowledgments and registration information for each mobile in a subsequent downlink message.

Schedule Message: The base station broadcasts a schedule message (SM) that contains the slot permissions for the subsequent data phase. Each permission identifies the mobile/VC combination that should transmit/receive in a given data slot. The data phase includes downlink transmissions from the base station, and uplink transmissions from the mobiles.

Each permission consists of a 2-bit type field, the MACid and VCid fields, and the length field spec-

Bit 0	*Bit 1*	*MACid*	*VCid*	*Description*
0	0	Receiving Mobile ID	VCid	BS to MH for specified VC
0	1	Sending Mobile ID	VCid	MH to BS for specified VC
1	0	Sending Mobile ID	VCid	Peer-to-peer and multicast support

Table 1: Packet types.

ifying the number of slots allocated to the VCid. The allocation information is grouped based on sender ID with a length field. This approach is attractive from an energy consumption perspective. It reduces the time the receiver has to be turned on to receive the schedule information. Table 1 lists the different types of permissions.

As wireless speeds increase to 20 Mbps and above, equalization is required at the receiver to establish bit synchronization. This synchronization may require around 25-30 bytes for each packet. Under such conditions, it will not be feasible to make allocations based on individual 53-byte cells, as is done with DQRUMA [12]. Therefore, our proposed protocol allocates clusters of slots to a sender.

Data Phases: Downlink transmission from the base station to the mobiles is scheduled considering the QoS requirements of the individual VCs. Likewise, the uplink slots are allocated using scheduling algorithm described below.

The definition of the protocol in terms of multiple phases in a frame is similar to other protocols proposed earlier [11, 13, 30, 31]. The new features of the proposed protocol are support for multiple traffic types, provision of per-VC queuing and scheduling, low power consideration, and provision of service quality to individual connections.

The protocol can be generalized to a system with multiple channels per cell. The channels can be either based on FDMA frequencies or CDMA codes. From the MAC point of view, the hybrid FDMA/TDMA and CDMA/TDMA systems have some similarities. However, the differences due to the underlying physical technology have to be carefully considered in fully defining the protocol. Fig. 4 shows an example EC-MAC frame structure for a hybrid CDMA/TDMA system. Following the frame synchronization are the simultaneous new-user and request/update phase. Some channels are reserved for new-user identification and others for currently registered mobiles. The downlink broadcast data, containing the schedule and other broadcast information, is sent on a single channel or replicated on all channels as required. The uplink data and downlink data phases can proceed in parallel provided that the receiver and transmitter can operate simultaneously on two different channels. If that is not possible, through appropriate scheduling, the BS can ensure that the receiver and transmitter of a given mobile are not used simultaneously. Note that this frame structure can also conceptually apply to a FDMA/TDMA system with multiple frequencies.

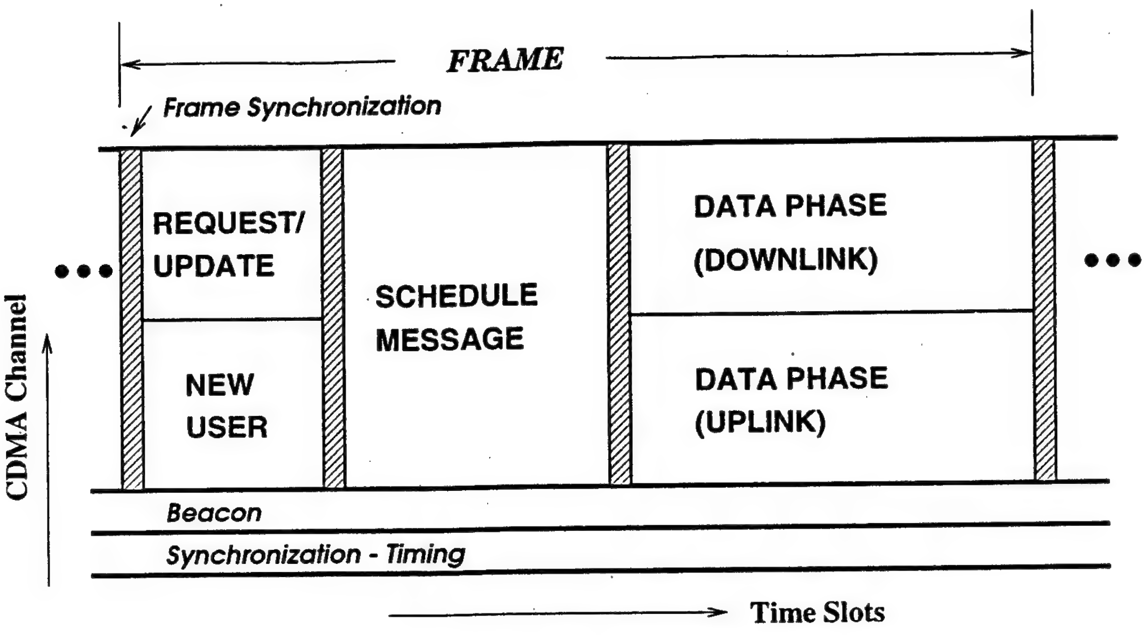


Figure 4: A hybrid CDMA/TDMA example.

4.2 Scheduling

The previous sections outlined the protocol definition and the multiple access mechanisms of EC-MAC. This section describes the scheduling algorithm associated with EC-MAC.

Many queuing and scheduling algorithms have been proposed for conventional wired ATM networks. An overview and comparison of some of the proposed algorithms can be found in [32,33]. Recently, algorithms have been proposed for the scheduling in wireless networks [14,15], but they do not address either the issue of low energy consumption or the diverse QoS requirements for heterogeneous traffic. The design of the proposed algorithm is described in the following sections.

Connection Admission Control (CAC): After a mobile is admitted to this cell – either locally generated or handed off, it may request bandwidth for several VCs as they are created. The CAC's goal is to maintain QoS for all existent VCs while admitting new VCs.

In [34], a measurement-based admission control for VBR video was proposed, where a new session is admitted based on measured utilization. In wireless environments, the measured utilization may not be the actual load of the system because some mobiles might be in error states. The measured utilization may fail to take those mobiles into account. Hence, this algorithm may admit more traffic than the system can afford.

In our algorithm, each VC sends the minimum guaranteed number of slots it needs as part of session set up. A counter is used to record the total number of slots which have been admitted. If the counter exceeds the number of slots in uplink data phase after adding this VC's request, the VC is rejected. Otherwise, it is admitted. The counter is incremented by the number of this VC's request slot(s). By this admission control algorithm, the total rate of all admitted VCs is always less or equal to the maximum capacity in data phase. Therefore, the QoS of existing VCs will not be affected by the newly admitted VC.

Scheduling: The proposed algorithm performs *coarse-grained scheduling* based on the frame structure of EC-MAC. Although many algorithms have been proposed for conventional wired ATM networks, most of them are based on packet-by-packet scheduling which are good for *fine-grained scheduling* only. For example, algorithms based on time-stamp such as Virtual Clock [35] and SCFQ [36] are not applicable for EC-MAC because they need to know the arrival time of each packet. Other type of algorithms such as HOL-EDD [37] might be modified for frame-based scheduling. Since it does not allow a session to be served at different rates at different times, a VBR video session cannot improve the delay performance without requesting the peak bandwidth. A *multirate* service algorithm was proposed to address the scheduling of VBR video [34]. However, this algorithm is fine-grained based on time-stamped priority. In addition, it was proposed for high-speed networks where errors are negligible.

The proposed algorithm is a *priority round robin with dynamic reservation update and error compensation* scheduling. The scheduler is currently defined to handle CBR, VBR, and UBR traffic. The scheduler gives higher priority to CBR and VBR traffic. These traffic sources can make requests for slot reservations that will be satisfied by the scheduler. UBR traffic, on the other hand, is treated with low priority and without reservation. Within the same traffic type, the different connections are treated using round robin mechanism.

The base station (BS) maintains two tables: *request table* and *allocation table*. The request table maintains the queue size of the virtual circuit of each mobile, the error state of the mobile, the number of requested reservations for CBR and VBR traffic, and the number of credits for UBR traffic. The purpose of the allocation table is to maintain the number of slots scheduled for each VC and each mobile. This table is essentially broadcast as the schedule to the mobiles. Based on this table, the base station allocates contiguous slots within a frame for each mobile.

The BS first allocates slots to CBR VCs which have been currently admitted. Because of the connection admission control (CAC) described above, CBR VCs that belong to mobiles in non-error (good) states are satisfied with their required rates. The CBR VCs are allocated X slots every Y frames, based on the traffic requirements. For instance, with a 12-ms TDMA frame, a 32-Kbps voice source is allocated one 48-byte slot per frame.

For sources with VBR traffic, the base station maintains the number of slots allocated in the previous frame. Let the current request of source i be C_i slots, and the allocation in previous frame be P_i slots. If $C_i < P_i$, C_i slots are allocated, and the remaining $P_i - C_i$ are released. If $C_i > P_i$, P_i slots are allocated in the first round. In the second round, extra slots available are evenly distributed among the VBR sources whose requests have not been fully satisfied in the first round.

Since there is correlation in a VBR video source, the reservation in current frame period represents the prediction for next frame. By the adjustment, the bandwidth allocation in each frame is different depending on the current traffic load and the number of packets generated by VBR sources. The reservation, hence, is updated dynamically in each frame for VBR traffic.

The BS then schedules UBR traffic after the scheduling of CBR and VBR. If the mobile is in error state, the base station adds credit(s) in the corresponding entry in request table. Otherwise, the base station either schedules slot(s) to this VC or schedules the aggregate credits this VC has until there is no more slot available. The reason for this credit is to ensure long-term fairness. This credit adjustment scheme is not applied to voice and video traffic since late packets will be dropped rather than be played back in such applications.

Contiguous Bandwidth Allocation: The allocation table is implemented as a two-dimensional array with one dimension for mobiles and the other dimension for the VCs in this mobile. The base station broadcasts the slot id and the number of slots for each VC by looking at the entry of each mobile. Therefore, all slots in same VC and all VCs in the same mobile can be transmitted together contiguously. By this allocation table, base station only needs to announce the allocation once in each TDMA frame. Mobiles also only need to turn on the transceiver once for all different types of packets.

Dealing with Errors: This section describes how the scheduling algorithm deals with bursty and location-dependent errors. At a time, only some of the mobiles may be capable to communicate with the base station – the others might be in error state. Since a mobile may encounter errors during any phase of the time frame, we discuss them individually as follows.

1. If a mobile is in error state during base station frame synchronization message (FSM) reception, it will not receive its transmission order. Thus, it will not send the request in the uplink of reservation phase, and the BS will mark the mobile as in error state. The scheduling algorithm might assign credits to the mobile depending on the traffic type.

   In case the mobile changes to good state any time after this phase, the mobile will not be able to transmit in the subsequent data phase. It could decide to receive broadcast packets.

2. If errors happen during the uplink of request/update phase, the BS will mark the mobile as in error state because it did not receive the transmission request. When the mobile sends request in the subsequent request/update phase, BS will mark the mobile as in good state. The situation is similar to the one above.

3. If errors happen while a mobile is receiving the schedule message, bandwidth that has been scheduled to this mobile will not be utilized. This loss is limited to only one data phase which is typically smaller than the average burst error length of $100ms$ [15]. The BS will mark the mobile as in error state when it does not receive this mobile's data during the scheduled uplink slots. The BS will mark the mobile as in good state when it receives the requests from mobiles in request/update phase again.

4. If errors happen during the downlink data phase or the mobile does not turn on receiver because of missed schedule message, the BS will hold packets until the corresponding mobile returns back to good state. Mobiles can acknowledge the packets they receive when they send requests in next request/update phase. Thus, the BS can know whether mobiles have received the downlink packets or not. The BS deletes packets from queues only after it receives acknowledgments.

5. If errors happen during the uplink of data phase, the BS will not receive the packets sent from the mobiles in error state. The BS acknowledges the packets it received in the next FSM. Mobiles delete packets from queues only after receiving acknowledgments or the deadline of real-time packets is expired. The BS will know the actual queue size of each VC and reschedule the packets when it receives the requests from mobiles in uplink request/update phase again.

This section described the mechanisms defined in EC-MAC to handle bursty and location-dependent errors during the various phases. The following section provides a simulation based performance analysis.

5 Performance Analysis

The following sections describe source traffic models, performance metrics studied, and simulation results for the protocol described above using realistic source traffic models for video, voice, and data services. A comparison of energy consumption for EC-MAC and other protocols is also provided.

The performance of the protocol has been studied through discrete-event simulation. Simulation results have been obtained using the stochastic self-driven discrete-event models, written in *C* with YACSIM [38]. YACSIM is a *C* based library of routines that provides discrete-event and random variate facilities. Steady state transaction times and utilization were measured.

5.1 Comparison of Energy Consumption

In the section, we compare the energy consumption of EC-MAC to a set of other access protocols that includes IEEE 802.11 [39], Packet Reservation Multiple Access (PRMA) [31], Multiservices Dynamic Reservation TDMA (MDR-TDMA) [13, 40], and Distributed-queuing Request Update Multiple Access (DQRUMA) [12]. Of these, 802.11 is designed primarily for data traffic, PRMA for voice and data traffic. The other three protocols are specifically designed to handle multimedia traffic. The results demonstrate the low-power characteristics of EC-MAC.

The comparison results presented here were obtained by mathematical analysis and are summarized from [29]. The complete details of the power consumption analysis may be found in [29] and have not been included here for the sake of brevity.

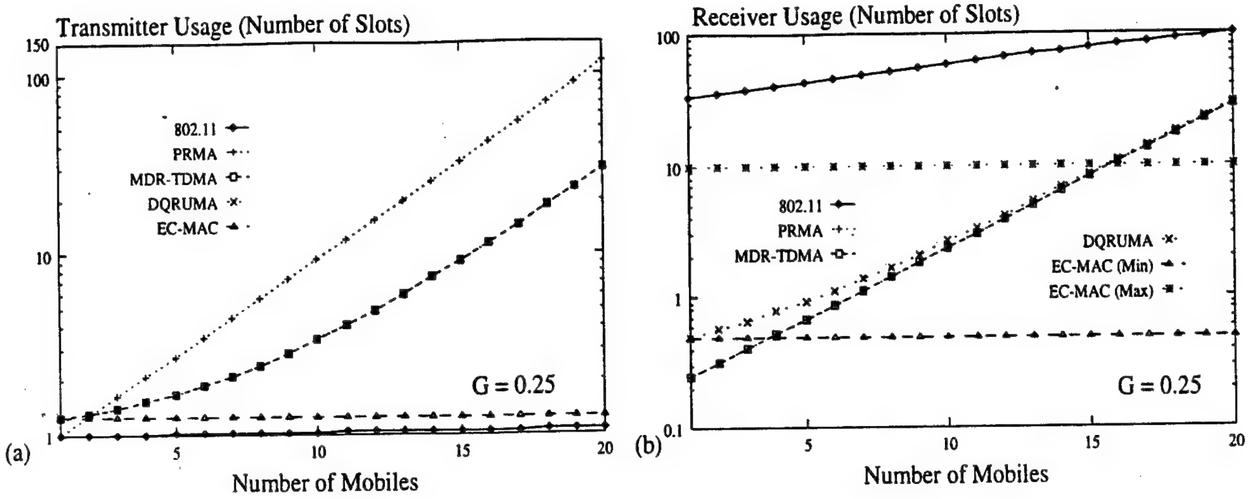


Figure 5: (a) Transmitter usage time , and (b) receiver usage time using Proxim's RangeLAN2 2.4 GHz radio, versus number of mobiles for transmitting a single packet. G is offered traffic load including newly generated and retransmitted packets.

The metrics compared are the transmitter and receiver usage time, and energy consumed for transmitting a single packet. In the figure, G denotes the offered traffic load which includes newly generated plus retransmitted packets. The results are obtained for a channel transmission rate of 2 Mbps. The packet length is 64 bytes. We assume that only 56 bytes are *useful* data after all coding schemes, header fields, error checksums, etc. are considered.

Figs. 5(a)-(b) show the transmitter and receiver usage times while transmitting a single packet. Although we show results for $G = 0.25$ only, the trends are similar for $G = 0.5$. But the increased load led to higher collisions for protocols based on Slotted Aloha. For 802.11, the mobile senses the medium before attempting to transmit. Collision occurs only when two or more mobiles choose the same slot in the contention window. The mobile transmits its packet after it captures the medium successfully in the contention window. Hence, the transmitter usage time in 802.11 is almost independent of the number of mobiles. However, the probability that the mobile under consideration contends successfully decreases as the traffic load increases. This results in increasing receiver usage time as the number of mobiles increases. Since the receiver is the most utilized resource in 802.11, the receiver usage in 802.11 is larger than others, while on the other hand, transmitter usage is much less than other protocols.

For PRMA, both receiver and transmitter need to be powered on in the slotted ALOHA contention mode. The transmitter is utilized for a packet transmission duration and the receiver is turned on to receive the acknowledgment. As the traffic load increases, the packet may suffer more collisions. Therefore, both the receiver and transmitter usage times increase. MDR-TDMA and DQRUMA also use slotted ALOHA to contend for a channel, but they employ a much shorter packet length. Hence, the two protocols have the same characteristics as PRMA does except that the time usage is less. In fig. 5 (a), MDR-TDMA and DQRUMA have the same transmitter usage time. Because reservation ALOHA is used in MDR-TDMA,

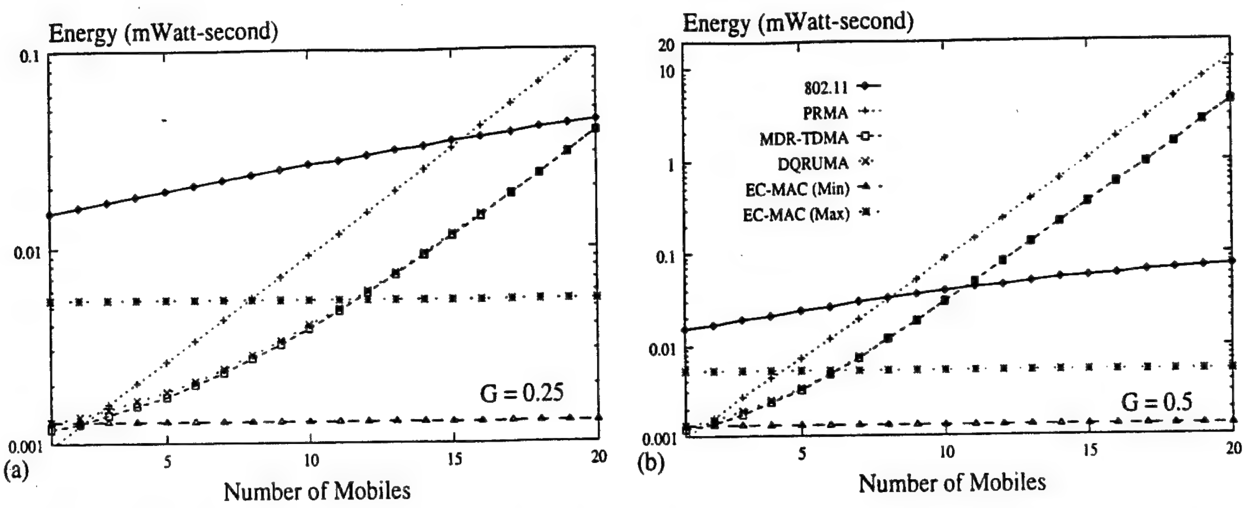


Figure 6: Energy spent per useful bit transmitted using Proxim's RangeLAN2 2.4 GHz radio, versus number of mobiles for transmitting a single packet. G is offered traffic load including newly generated and retransmitted packets. The figures are plotted for $G \in \{0.25, 0.5\}$.

packets in MDR-TDMA know the slot to transmit after the initial contention. In DQRUMA, however, the mobile needs to listen to transmission permissions explicitly for every slot.

Both the receiver and transmitter usage time remain constant in EC-MAC in fig. 5. Figs. 5(a)-(b) indicate that transmitter usage time is quite small in comparison to other protocols. It is very close to 802.11 when the load is heavy. Depending on how long the mobile listens to the transmission order and schedule message, the receiver usage time may be greater or less than other protocols. The receiver usage time in EC-MAC, however, is independent of the traffic load.

Figs. 6(a)-(b) provide an approximate comparison of energy spent per useful bit transmitted, while transmitting a single packet using Proxim's radio card. Note that the Proxim radio has been used merely to obtain typical power consumption values. It does not imply that all these access protocols can be implemented on a Proxim card. The results should therefore be construed as merely indicative of the performance trends.

Since MDR-TDMA and DQRUMA use a short packet for contention, they consume less energy than PRMA does. IEEE 802.11 senses the channel before transmission, reducing collision. However, it may need to sense several slots before it captures the medium. Therefore, 802.11 consumes more energy than PRMA, MDR-TDMA and DQRUMA do in lightly-loaded systems. On the other hand, during heavy system traffic there might be too many contentions for slotted ALOHA. We can see that 802.11 performs better than MDR-TDMA and DQRUMA when there are around 10 mobiles in fig. 6(b), and that the energy consumption of EC-MAC is independent of the traffic load and number of mobiles. In fact, we see that even the upper bound of energy consumption of the EC-MAC protocol can be significantly less than other protocols for heavily-loaded systems.

This section compared the energy consumption performance of EC-MAC to the other protocols. The sub-

Items	Value
Channel rate	10176 Kbps
TDMA frame length	12 ms
Number of slots (cells) per TDMA frame	288 slots
Number of slots (cells) per Data Phase	266 slots
Voice/Data/Video slot size	53 bytes
Average length of a voice call	3 min
Percentage of talkspurts for a voice call	36%
Percentage of silence gaps for a voice call	64%
Average length of a talkspurt	1.00 sec
Average length of a silent gap	1.35 sec
Speech coding rate	32 Kbps
Average length of a video call	5 min
Maximum video cell delay	144 ms
Hurst parameter (index of self-similarity)	0.9

Table 2: System parameters used for simulation.

sequent sections provide a detailed performance analysis of EC-MAC using voice, video, and data traffic models with diverse QoS requirements.

5.2 Source Models

The simulation results presented here consider three types of traffic – one each for CBR, VBR, and UBR category. Voice is modeled as a two-phase process with talkspurts and silent gaps [16]. Typically, such modeling classifies voice as VBR. We consider that the voice source generates a continuous bit-stream during talkspurts and is therefore classified as a CBR source in our scheduling. Video is considered as an example of a VBR source with variable number of cells per frame. Data generated by applications such as ftp, http and email is considered as an example of UBR traffic.

In simulation, each mobile terminal is capable of generating three different types of traffic: data, voice, and video. An idle mobile generates new voice calls and video calls with rates of λ_s, and λ_v, respectively. Data traffic is modeled as self-similar traffic with Hurst parameter of 0.9 (described below). The following paragraphs present the simulation models for data, voice, video, and error, respectively. The system parameters are summarized in table 2.

Data Model: Recently, extensive studies show that data traffic is self-similar in nature, and the traditional Poisson process cannot capture this fractal-like behavior [19]. The difference between self-similar and traditional models is that the self-similar model is long-range dependent, i.e., bursty over a wide range of time scales. Self-similar model shows that the traffic has similar statistical properties at a range of time

Name	Mean rate	Max rate	Name	Mean rate	Max rate
Car phone 1	50.03	464.00	Car phone 2	40.56	409.40
Claire 1	33.79	658.00	Claire 2	23.20	550.60
Foreman 1	44.11	404.40	Foreman 2	31.82	368.80
Grandma 1	13.98	481.94	Grandma 2	8.33	370.20
Mother & daughter 1	26.18	525.82	Mother & daughter 2	15.72	391.40
Miss America 1	37.19	446.00	Miss America 2	23.87	392.60
Salesman 1	17.22	506.60	Salesman 2	12.69	464.00
Suzie 1	31.07	322.80	Suzie 2	22.88	285.20
Trevor 1	30.57	406.80	Trevor 2	24.95	366.40

Table 3: Video bit rates (in Kbps)

scales: milliseconds, seconds, minutes, and hours. Long-range dependent traffic (fractional Gaussian noise) can be obtained by the superposition of many ON/OFF sources in which the ON and OFF periods have a Pareto type distribution with infinite variance [20]. In simulation, we use the strictly alternating ON/OFF sources with the same α-value for the Pareto distribution. The α value equals 1.2 which corresponds to the estimated *Hurst parameter*, the index of self-similarity, of $H = 0.9$ [20].

Voice Model: A voice source is modeled as a two-state Markov process representing a source with a *slow speech activity detector* (SAD) [16]. The probability that a principal talkspurt with mean duration t_1 seconds ends in a frame of duration τ is $\gamma = 1 - exp(-\tau/t_1)$. The probability that a silent gap with mean duration t_2 seconds ends in a frame of duration τ is $\sigma = 1 - exp(-\tau/t_2)$. Here, γ is the probability that a source makes a transition from talkspurt state to silent state, and σ is the probability that the source makes a transition from silent state to talkspurt state.

Measured values for t_1 and t_2 are 1.00 sec and 1.35 sec [16], each with exponential distribution. This results in an average of 36% talkspurts and 64% silence gaps for each voice conversation. A voice cell is dropped if not transmitted after 36 ms. When a new voice cell arrives at a full queue, the *first* cell in the voice queue will be dropped.

Video Model: In simulation, we used the real trace data from several H.263 video sources shown in table 3 [41]. H.263 [17, 18] targets the transmission of video telephony at data rates less than 64 Kbps which makes it suitable for wireless communications. Each video is coded by two different schemes. The first one has I and P frames only. The second one adds some other options such as PB-frames and advanced prediction mode. Each video runs for around 30 seconds. The frame rate of *Grandma 1* and *Mother & daughter 1* is 30 fps. All others have frame rate of 25 fps. Table 3 shows the bit rates of the video sources. The table shows that the second coding scheme has lower bandwidth requirements for all the sources.

For a TDMA frame of length 12 ms (as used in the simulation), the mean number of video packets is around

1 ATM cell per TDMA frame and the maximum is 21 ATM cells per TDMA frame. In the simulation, we assume that the length of a video session is exponentially distributed with mean time of 5 minutes. This is achieved by randomly selecting different videos (since each video trace only lasts above 30 seconds).

Error Model: In high-speed networks based on fiber optics, errors are rare and random in nature. In wireless networks, errors are bursty and location-dependent. A finite state Markovian model can be effectively used to characterize bit error patterns on RF channels. In [15], authors observed that mean residency time in states with high BER (bit error rate) is longer than a single packet transmission time. Therefore, a single packet loss would be followed by many back-to-back packet losses. The authors also defined a two-state Markov model to characterize the packet loss. The channel may be in G (good) or B (bad) states. The corresponding packet loss probabilities are p^G and p^B in state G and state B respectively, where $p^B \gg p^G$. The time spent in the Good and Bad periods are g and b respectively, each exponentially distributed. In this model, packet error rate, e, equals:

$$e = \frac{g\, p^G + b\, p^B}{g + b}$$

In experiments, the mean burst length of b was set to $100ms$, $p^B = 0.8$, and $p^G = 0$ [15]. In the simulation, $e = 10^{-3}$ and $e = 10^{-5}$ have been used, and the corresponding g was obtained by the above equation.

Other models for a Rayleigh fading channel have been studied as in [42]. Here, it has been shown that a first-order Markov model is an adequate approximation for a Rayleigh fading channel. The performance of an access protocol with such a model with capture has been studied in [43]. We are investigating these other models too for incorporation in our studies.

5.3 Performance metrics studied

The focus of the study is to understand what kind of service quality is provided by the protocols with an increase in the number of mobiles supported. To this end, we define the following QoS parameters:

Voice-cell dropped rate: The voice-cell dropped rate is defined as the ratio of *the number of voice cells dropped* to *the total number of voice cells generated.* It has been suggested in [31] that it should not exceed 1%, since distortion will be perceptible otherwise.

Voice-call dropped rate: The voice-call dropped rate is defined as the ratio of *the number of voice calls dropped* to *the total number of voice calls generated.* The acceptable value depends on the requirement of each system. In [40], the authors have considered 1% voice call blocking probability.

Video-cell dropped rate: The video-cell dropped rate is defined as the ratio of *the number of video cells dropped* to *the total number of video cells generated.* The desired value depends on different coding algorithms and different type of services. For non-layered MPEG-2 coding, [44] indicates that the packet loss ratios of 10^{-3} or greater for ATM cells are generally unacceptable. However, [45] shows that the quality could be improved when the cell loss rate is greater than 10^{-3} by macroblock re-synchronization technique.

Video-call dropped rate: The video-call dropped rate is defined as the ratio of *the number of video calls dropped* to *the total number of voice calls generated.* The acceptable value depends on the requirement of each system. In order to maintain the quality of existent sessions, the CAC can restrict the number of video sessions because they generally require much more bandwidth.

Average Data Cell Delay: The average data cell delay is defined as *the time a data cell transmitted* minus *the time a data cell generated.* The desired value depends on the characteristics of different type of data services, such as file transfer or e-mail, etc.

Channel Utilization: The channel utilization is defined as the ratio of *the number of slots used for transmission* to *the total number of slots available.* Since the study is focused on how well a protocol can schedule mobiles for uplink transmission, we consider the uplink channel utilization only.

5.4 Simulation Results

The numerical results presented here study the maximum number of mobiles that can be accommodated with the desired QoS for voice, data, and video traffic. A channel rate of 10 Mbps has been considered. Each uplink and downlink in data phase is around 4.7 Mbps. The figures are plotted with offered load of 25% and 50% per mobile. Data traffic is modeled as self-similar traffic with Hurst parameter of 0.9 [20]. When load is 50%, the inter-arrival times of voice calls ($1/\lambda_s$) and video calls ($1/\lambda_v$) are 180 sec and 300 sec, respectively. The average length of a voice call is 3 minutes, so the voice traffic load is $180/(180+180) = 50\%$. The average length of a video call is 5 minutes, so the video traffic load is $300/(300+300) = 50\%$ also. The *packet error rates* are 10^{-3} and 10^{-5}.

Voice-call dropped rate is considered first in fig. 7(a). As expected, less voice calls are dropped if less mobiles contend in the system. With the same number of mobiles, traffic load of 50% leads to higher dropped rate than traffic load of 25%. Fig. 7(a) also shows that the major factor for call dropped rate is traffic load rather than the error rate, as expected. When the load is 50% and the number of mobiles is greater than 80, the voice call dropped rate increases rapidly. For the load of 25%, voice call dropped rate is acceptable even when the number of mobiles is 160 regardless of the error rate.

Once a voice call is admitted, fig. 7(b) indicates that the voice-cell dropped rate is independent of traffic load. The cell dropped rate is very close to the error rate regardless of the number of mobiles and the load offered by each mobile. The CAC algorithm restricts the number of connections to maintain the QoS of admitted connections. The CAC and scheduler cooperate with each other like this: CAC deals with traffic load and scheduler deals with QoS requirements. Although cells are still dropped, that is the *de facto* nature of wireless channel due to errors. Fig. 7(b) shows that the voice-cell dropped rate is only slightly higher than

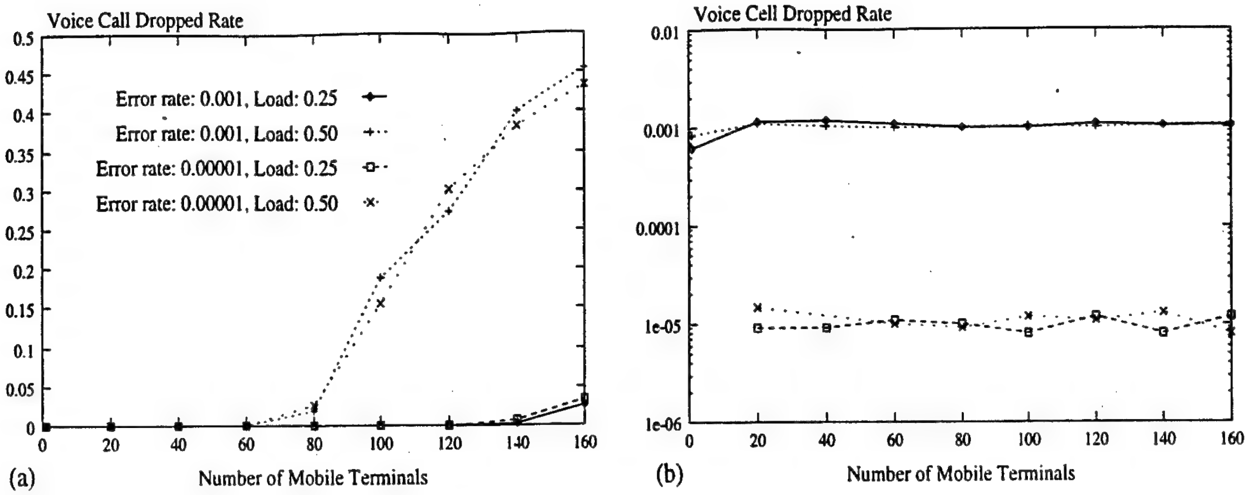


Figure 7: Voice (a) call dropped rate, (b) cell dropped rate. (Channel rate: 10 Mbps).

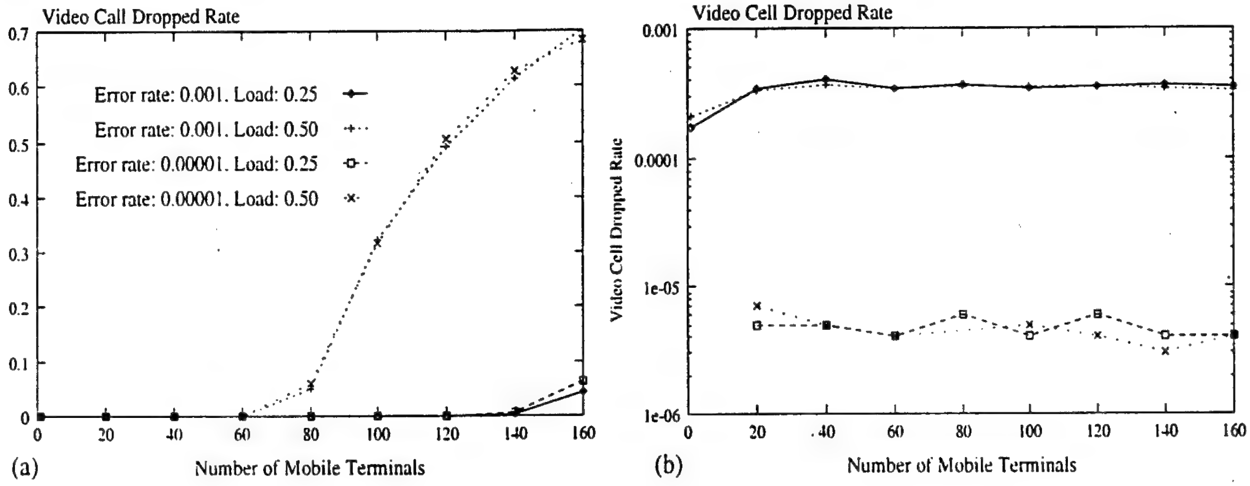


Figure 8: Video (a) call dropped rate, and (b) cell dropped rate. (Channel rate: 10 Mbps).

the error rate even when there are many mobiles in the system. For 10^{-5} error rate, the voice-cell dropped rate is 0 when the number of mobiles is less than 20. Fig. 7(b), therefore, shows the number of mobiles starting from 20.

Figs. 8(a) and (b) examine the video-call and video-cell dropped rates. As discussed above, call dropped rate is determined mainly by the offered traffic load. With CAC, a video call sends the minimum guaranteed rate it needs. Based on this information, CAC decides to admit or reject this call. If a video call requests the maximum rate it needs in CAC, there will be no dropped cell ideally. However, it is wasteful to decide on admission control based on maximum bandwidth requirement. If it requests a mean rate in CAC, many other sessions can be admitted but the cell dropped rate may be unacceptable. For a H.263 video with mean rate of 1 ATM cells per TDMA frame and peak rate of 21 cells per TDMA frame, figs. 8(a) and (b) show the results when each video sets 2 cells as the minimum guaranteed rate. Fig 8(b) indicates that the video-cell

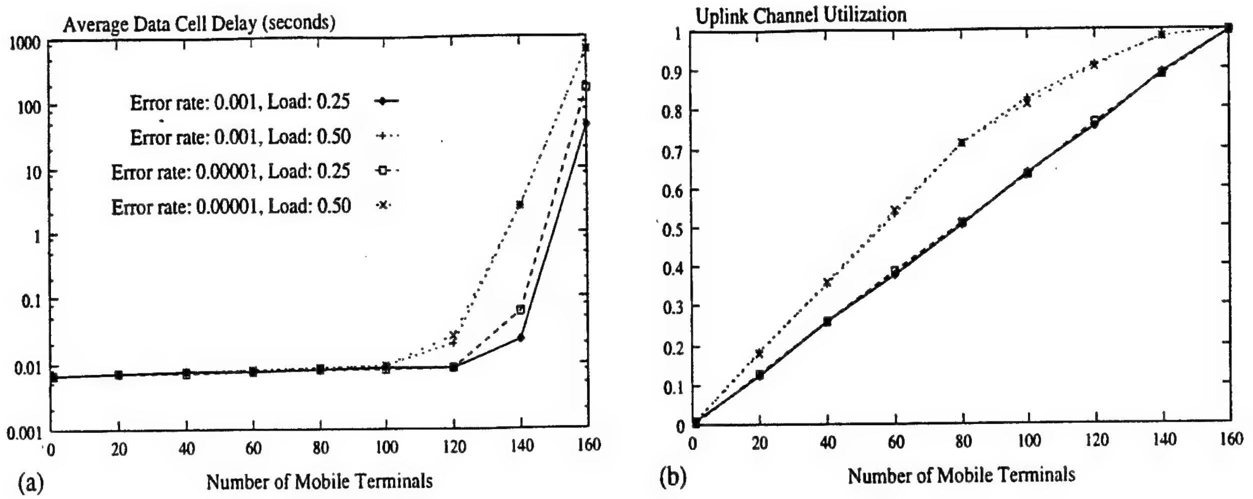


Figure 9: (a) Data cell delay, and (b) Channel Utilization.(Channel rate: 10 Mbps)

dropped rate is smaller than the error rate although the request rate set in CAC is still much less than the peak video rate. This indicates that our *dynamic reservation update* scheme can get a good multiplexing gain. Fig. 8(a) also shows our algorithm can support 80 video sessions when load equals 0.5. More than 160 video sessions can be accepted when load equals 0.25 if the required call dropped rate is set to 1%.

Fig. 9(a) compares the data cell delay. Data traffic is transmitted when there are no other voice or video traffic pending. Although data is with lower priority and without any reservation, it still gets chances to transmit when some voice or video sessions are in error state, or when VBR video sessions generate less traffic. As expected, the data cell delay increases when the number of mobiles increases. Fig. 9(a) shows that the higher load generally has higher data delay. Channel error rate, however, will not affect data delay too much. This is because the scheduling algorithm credits the error mobiles after they change back to good state.

Fig. 9(b) examines the uplink channel utilization. When the traffic load is higher, overall utilization is higher. The error rate has a little performance difference for utilization since the error rates are small comparing to the total bandwidth. The channel utilization increases as the number of mobiles increases.

6 Summary

This paper describes an access protocol for wireless and mobile ATM networks. The goals of the access protocol are to conserve battery power, to support multiple traffic classes, and to provide different levels of service quality for bandwidth allocation. The protocol is based on a combination of reservation and scheduling mechanisms. The protocol architecture and the design decisions have been outlined in this paper. Performance analysis based on discrete simulation has been provided which studies various quality-of-service parameters with varying number of mobiles in a cell. The energy consumption comparison of EC-MAC to other protocols including IEEE 802.11 standard has been provided. The comparison demonstrates that EC-MAC has better energy consumption characteristics because of its collision-less nature. A detailed performance analysis with voice, video and data traffic models was presented. This analysis gives an indication of the number of the mobiles that can be supported. By the scheduling algorithm associated with this protocol, we show that EC-MAC, in addition to low-power consumption, can achieve high channel utilization, low packet delay, and meet the QoS requirements for multimedia traffic.

To summarize the performance, we show that the protocol can support between 80 and 160 mobiles with low and intermediate loads. Quality-of-service is maintained by using the scheduling algorithm in conjunction with the admission control algorithm. A number of other performance issues that need to further investigated include: (i) how many slots are totally allocated to VBR and CBR sources, (ii) the performance benefits achieved by using contiguous allocation, and (iii) the delay implications due to contiguous allocation. This is the subject of further study.

References

[1] D. C. Cox, "Wireless personal communications: What is it?," *IEEE Personal Communications*, vol. 2, pp. 20–35, Apr. 1995.

[2] K. Pahlavan and A. H. Levesque, "Wireless data communications," *Proceedings of the IEEE*, vol. 82, pp. 1398–1430, Sept. 1994.

[3] M. D. Prycker, *Asynchronous Transfer Mode: Solution for Broadband ISDN*. Prentice Hall, 3 ed., 1995.

[4] M. Garrett, "A service architecture for ATM: from applications to scheduling," *IEEE Network*. pp. 6–14, May/June 1996.

[5] M. Naghshineh (Guest Ed.), "Special issue on Wireless ATM," *IEEE Personal Communications*, vol. 3, Aug. 1996.

[6] M. J. Karol and K. Sohraby (Guest Ed.), "Special issue on Wireless ATM," *ACM/Baltzer Mobile Networks and Applications*, vol. 1, Dec. 1996.

[7] T. R. Hsing, D. C. Cox, L. F. Chang, and T. Van Landegem (Guest Ed.), "Special issue on Wireless ATM," *IEEE Journal on Selected Areas in Communications*, vol. 15, Jan. 1997.

[8] L. Dellaverson, "Wireless ATM Working Group – Charter and Overview." Charter for ATM Forum's Wireless ATM Working Group, Sept. 1996.

[9] S. Singh, "Quality of Service guarantees in mobile computing," *Computer Communications*, vol. 19, pp. 359–371, Apr. 1996.

[10] C.-S. Chang, K.-C. Chen, M.-Y. You, and J.-F. Chang, "Guaranteed quality-of-service wireless access to ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 106–118, Jan. 1997.

[11] K. S. Natarajan, "A hybrid medium access control protocol for wireless LANs," in *Proc. 1992 IEEE International Conference on Selected Topics in Wireless Communications*, (Vancouver, B.C., Canada), June 1992.

[12] M. J. Karol, Z. Liu, and K. Y. Eng, "An efficient demand-assignment multiple access protocol for wireless packet (ATM) networks," *ACM/Baltzer Wireless Networks*, vol. 1, no. 3, pp. 267–279, 1995.

[13] D. Raychaudhuri, L. J. French, R. J. Siracusa, S. K. Biswas, R. Yuan, P. Narasimhan, and C. A. Johnston, "WATMnet: A prototype wireless ATM system for multimedia personal communication," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 83–95, Jan. 1997.

[14] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," in *Proc. ACM SIGCOMM*, (Palais des Festivals, Cannes, France), Sept. 1997.

[15] P. Bhagwat, P. Bhattacharya, A. Krishna, and S. K. Tripathi, "Enhancing throughput over wireless LANs using channel state dependent packet scheduling," in *Proc. IEEE INFOCOM*, (San Francisco, CA), pp. 1133–1140, Apr. 1996.

[16] D. J. Goodman and S. X. Wei, "Efficiency of packet reservation multiple access," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 170–176, Feb. 1991.

[17] ITU-T Rec. H.263, "Video coding for low bit rate communication," Mar. 1996.

[18] K. Rijkse, "H.263: video coding for low-bit-rate communication," *IEEE Communications Magazine*, vol. 43, pp. 42–45, Dec. 1996.

[19] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, pp. 1–15, Feb. 1994.

[20] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 71–86, Feb. 1997.

[21] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. B. Srivastava, and J. A. Trotter, "SWAN: A mobile multimedia wireless network," *IEEE Personal Communications*, vol. 3, pp. 18–33, Apr. 1996.

[22] M. Stemm and R. H. Katz, "Measuring and reducing energy consumption of network interfaces in hand-held devices," *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Science*, Aug. 1997.

[23] M. Zorzi and R. R. Rao, "Error control and energy consumption in communications for nomadic computing," *IEEE Transactions on Computers*, vol. 46, pp. 279–289, Mar. 1997.

[24] M. Zorzi and R. Rao, "Energy constrained error control for wireless channels," *IEEE Personal Communications*, Dec. 1997.

[25] P. Lettieri, C. Fragouli, and M. B. Srivastava, "Low power error control for wireless links," in *Proc. ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, (Budapest, Hungary), Sept. 1997.

[26] K. M. Sivalingam, M. B. Srivastava, P. Agrawal, and J.-C. Chen, "Low-power access protocols based on scheduling for wireless and mobile ATM networks," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, (San Diego, CA), pp. 429–433, Oct. 1997.

[27] D. Petras and A. Kramling, "MAC protocol with polling and fast collision resolution for an ATM air interface," in *Proc. IEEE ATM Workshop*, (San Francisco, CA), Aug. 1996.

[28] ETSI-RES10, "High performance radio local area network (HIPERLAN)." ETS 300, Feb. 1997.

[29] J.-C. Chen, K. M. Sivalingam, P. Agrawal, and S. Kishore, "A comparison of MAC protocols for wireless local networks based on battery power consumption," in *Proc. IEEE INFOCOM*, (San Francisco, CA), Apr. 1998. To appear.

[30] G. Bianchi, F. Borgonovo, L. Fratta, L. Musumeci, and M. Zorzi, "C-PRMA: A centralized packet reservation multiple access for local wireless communications," *IEEE Transactions on Vehicular Technology*, vol. 46, pp. 422–436, May 1997.

[31] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," *IEEE Transactions on Communications*, vol. 37, pp. 885–890, Aug. 1989.

[32] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, vol. 83. Oct. 1995.

[33] A. Varma and D. Stiliadis, "Hardware implementation of fair queuing algorithms for asynchronous transfer mode networks," *IEEE Communications Magazine*, vol. 35, pp. 54–68, Dec. 1997.

[34] D. Saha, S. Mukherjee, and S. K. Tripathi, "Multirate scheduling of VBR video traffic in ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1132–1147, Aug. 1997.

[35] L. Zhang, "VirtualClock: a new traffic control algorithm for packet switching networks," *ACM Transactions on Computer Systems*, vol. 9, pp. 101–124, May 1991.

[36] S. Golestani, "A self-clocked fair queueing scheme for broadband applications," in *Proc. IEEE INFOCOM*, (Toronto, Ont., Canada), pp. 636–646, June 1994.

[37] M. Vishnu and J. W. Mark, "HOL-EDD: A flexible service scheduling scheme for ATM networks," in *Proc. IEEE INFOCOM*, (San Francisco, CA), pp. 647–654, Apr. 1996.

[38] J. R. Jump, *YACSIM Reference Manual*. Rice University, Department of Electrical and Computer Engineering, 2.1 ed., Mar. 1993.

[39] IEEE, "Wireless LAN medium access control (MAC) and physical layer (PHY) Spec." P802.11/D5, Draft Standard IEEE 802.11, May 1996.

[40] D. Raychaudhuri and N. D. Wilson, "ATM-based transport architecture for multi-services wireless personal communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 1401–1414, Oct. 1994.

[41] "Digital video coding at Telenor R&D." http://www.fou.telenor.no/brukere/DVC/.

[42] M. Zorzi, R. Rao, and L. B. Milstein, "On the accuracy of a first-order Markov model for data transmission on fading channels," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, pp. 211–215, Nov. 1995.

[43] A. Chockalingam, M. Zorzi, L. B. Milstein, and P. Venkataram, "Performance of a wireless access protocol on correlated Rayleigh fading channels with capture," *IEEE Transactions on Communications*, 1998. To Appear.

[44] R. Aravind, M. R. Civanlar, and A. R. Reibman, "Packet loss resilience of MPEG-2 scalable video coding algorithms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 426–435, Oct. 1996.

[45] J. Zhang, M. R. Frater, J. F. Arnold, and T. M. Percival, "MPEG 2 video services for wireless ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 119–128, Jan. 1997.

UNDER REVIEW WITH IEEE JSAC, 1997.

Dynamic resource allocation schemes during handoff for mobile multimedia wireless networks

Parameswaran Ramanathan[1], Krishna M. Sivalingam[2,*], Prathima Agrawal[3], and Shalinee Kishore[4]

[1] Dept. of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706
[2] School of Electrical Engg. & Computer Science, Washington State University, Pullman, WA 99164
[3] AT&T Labs, Whippany, NJ 07981
[4] WINLAB, Rutgers University, Piscataway, NJ 08855

Abstract

User mobility management is one of the important components of mobile multimedia systems. In a cell-based network, a mobile should be able to seamlessly obtain transmission resources after handoff to a new basestation. This is essential for both service continuity and quality of service assurance. In this paper, we present strategies for accommodating continuous service to mobile users through estimating resource requirements of potential handoff connections. A diverse mix of heterogeneous traffic with diverse resource requirements is considered. We investigate static and dynamic resource allocation schemes. The dynamic scheme probabilistically estimates the potential number of connections that will be handed off from neighboring cells, for each class of traffic. The performance of these strategies in terms of connection blocking probabilities for handoff and local new connection requests are evaluated. The performance is also compared to a scheme previously proposed in [1]. The results indicate that using dynamic estimation and allocation, we can significantly reduce the dropping probability for handoff connections.

[*] Part of the research was supported by Air Force of Scientific Research grant F-49620-97-1-0471, and by AT&T Labs, Whippany, NJ. The authors can be reached at parmesh@ece.wisc.edu, krishna@eecs.wsu.edu, pa@research.att.com, and kishore@ece.rutgers.edu.

1 Introduction

A wireless network is typically organized into geographical regions called *cells* [2]. The mobile users in a cell are served by a *basestation*. Before a mobile user can communicate with other user(s) in the network, a *connection* must usually be established between the users. The establishment and maintenance of a connection in a wireless network is the responsibility of the basestation. To establish a connection, a mobile user must first specify its traffic characteristics and Quality of Service (QoS) needs. This specification may be either implicit or explicit depending on the type of services provided by the network. For example, in a cellular phone network, the traffic characteristics and the QoS needs of voice connections are known apriori to the basestation and, therefore they are usually implicit in a connection request. However, in the future, wireless networks will have to provide support for multimedia services where, the traffic characteristics and the QoS needs of a connection may not be known apriori to the basestation. In this case, the mobile user must explicitly specify the traffic characteristics and QoS needs as part of the connection request. Wireless ATM networks are an example of such a network [3–5]. In either case, the basestation determines whether it can meet the requested QoS needs and, if possible, establish a connection.

When a user moves from one cell to another, the basestation in the new cell must take responsibility for all the previously established connections. A significant part of this responsibility involves allocating sufficient resources in the cell to maintain the QoS needs of the established connection(s). If sufficient resources are not allocated, the QoS needs may not be met, which in turn may result in premature termination of the connection. Since premature termination of established connections is usually more objectionable than rejection of a new connection request, it is widely believed that a wireless network must give higher priority to the handoff connection requests as compared to new connection requests. Many different admission control strategies have been discussed in literature to provide priorities to handoff requests without significantly jeopardizing the new connection requests [1,6–10].

The basic idea of these admission control strategies is to apriori reserve resources in each cell to deal with handoff requests. In conventional cellular networks, where the traffic and QoS needs of all connections are the same, the reservation of resources typically occurs in the form of "guard channels", where a new connection request is established if and only if the total available channels or capacity is greater than a pre-determined threshold [1,7–10]. The strategies differ in how the number of guard channels (i.e., the threshold) is chosen by a basestation.

A simple strategy is to reserve a fixed percentage of the basestation's capacity for handoff connections. If this percentage is high, adequate capacity will most likely be available to maintain the QoS needs of handoff connections, but, at the expense of rejecting new connections. The advantage of this strategy is, of course, its simplicity because there is no need for exchange of control information between the basestations. A more involved, but possibly, better strategy is for each basestation to dynamically adapt the capacity reserved for dealing with handoff requests based on the current connections in the neighboring cells. This will enable the basestation to reserve an approximately the actual resources needed for handoff requests and thereby accept more new connection requests as compared to in a fixed scheme. Such dynamic strategies are proposed and evaluated in [1, 9].

In [9], Naghshineh and Schwartz develop a theoretical model to compute the resource requirements for handoff requests so as to maintain a target handoff blocking probability. This is the probability of not having adequate capacity to allocate to a handoff request. Their model assumes that all connection requests are identical and the analysis is carried out for a simple three cell configuration under stationary traffic conditions. In [1], Yu and Leung also propose a technique to compute the capacity to be reserved for handoff requests so as to either strictly or loosely maintain the handoff blocking probability within a specified target. They also simulate a more realistic multicell wireless network and compare the performance of their strategy with that of a static strategy. To estimate the future probability of blocking, they assume Poisson arrival of new connection requests, Poisson arrival of handoff connection requests, exponential connection duration, and exponential channel holding time. Note that, channel holding time for a connection in a cell depends on the unencumbered cell residence time (i.e., cell residence time if the connection is of an infinite duration) and the remaining connection duration. In practice, unencumbered cell residence time may not be exponentially distributed [11], in which case, the strategy proposed in [1] will not be theoretically valid. Also, as in [9], Yu and Leung's model assumes that all connection requests are identical, which is not valid if multimedia services are to be supported by the wireless network.

In contrast, in this paper, we consider a wireless network supporting diverse traffic characteristics of voice, data, and video applications. Since the connections can now differ in the amount of resources (say bandwidth) required to meet their QoS needs, the question is how should a basestation dynamically adapt the amount of resources reserved for dealing with handoff requests. The strategy proposed in this paper is an approximation of the ideal strategy described below.

Consider an ideal wireless network in which each basestation knows the exact arrival times and resource requirements of all future handoff requests, and the completion times and the cell residence times of connections presently in its cell. Now suppose a new connection request comes into a basestation at time t and let T be the amount of time this connection will spend in the cell. Further suppose that, the objective is to accept all handoff requests. Then, basestation must accept the new connection request, if and only if, the additional resources needed to accept all incoming handoff requests in the interval $(t, t+T)$ plus the resources needed to support the new request is less than the amount of resources available at time t. This strategy is ideal because a basestation can only estimate the arrival times of handoff requests, the resource requirements of handoff requests, and the residence time of connections in the cell. Therefore, in the proposed approach, a basestation first estimates T, the expected cell residence time of the new connection request and the expected maximum additional resources needed to accept all incoming handoff requests in the interval $(t, t+T)$. If the estimated maximum additional resources needed to deal with handoffs plus the resources needed to support the new connection requests is less than the resource available at time t, then the new request is accepted.

The blocking probabilities for handoff and new connection requests in the proposed strategy is evaluated using a discrete-event simulator of a cellular network in a metropolitan area. The simulator also implements an extended version of the strategy proposed in [1] and two static schemes. The strategy in [1] is extended to deal with connection requests with different traffic characteristics. A comparison of the blocking probabilities show that the handoff blocking probability is among the smallest for the proposed scheme in different network types and traffic scenario. The traffic scenarios simulated include the morning rush-hour situation, evening rush-hour situation, and the mid-day high load situation. The simulation also shows that an extended version of the strategy in [1] does not always perform better than a static scheme when connections with diverse traffic requirements are present.

The rest of the paper is organized as follows. Section 2 presents our assumed model of the wireless network and reviews details of related strategies from literature. Section 3 describes the proposed strategy. Section 4 presents results of an empirical evaluation of the proposed strategy. Section 5 provides the summary and conclusions.

2 System Model and Related Work

A basestation in a cellular network may receive new connection requests from mobile users within its cell as well as handoff requests from mobile users in the neighboring cells. As part of a connection request, a mobile user promises to adhere to certain traffic characteristics and in return seeks some quality of service (QoS) guarantees from the network. The connections may differ in the traffic characteristics (constant bit rate, variable bit rate) and the desired QoS guarantees (e.g., delay bound, loss bound, throughput). In this paper, we assume that the promised traffic characteristics and the desired QoS guarantees can together be represented by a single number called the *effective bandwidth* of the connection.

Techniques for computing the effective bandwidth for different traffic characteristics and QoS requirements have been discussed elsewhere in literature [12–15], and is not the focus of this paper. For example, given a traffic envelope (i.e., a bound on the number of bytes generated by the user in any given time interval) and a desired delay bound, Le Boudec discusses an approach for computing the effective bandwidth which completely characterizes the envelope and the delay requirement [15]. Similarly, given stochastic characteristics of the traffic, the buffer size at a network element, and a desired bound on probability of packet loss, many different techniques have been proposed to compute the equivalent effective bandwidth [12–14].

Given the effective bandwidths of all the active connections in a cell and the effective bandwidth of a new connection request, the QoS requirements of all connections can be guaranteed if the sum of the effective bandwidths including the new request is less than or equal to the capacity of the cell. If this simple admission control criterion is used to accept both new and handoff connection requests, then the blocking probability for both types of requests will be the same. However, since it is desirable to have smaller blocking probabilities for handoff requests, the proposed strategy is based on the admission control scheme shown in Figure 1.

In Figure 1 and in rest of this paper, we assume that the connection requests in the network belong to one of M diverse classes. The classes correspond to different multimedia applications like voice, data, and video which are expected to run on future wireless networks. From the point of view of the wireless

```
Admission Control
    If incoming request belongs to class $\tau$
    If incoming request is a handoff then
        If available bandwidth $> \Delta_{h,\tau} + \phi_\tau$ then
            Accept
        Else
            Reject
        End
    End
    Else /* it is a new connection request */
        If available bandwidth $> \Delta_{n,\tau} + \phi_\tau$ then
            Accept
        Else
            Reject
        End
    End
```

Figure 1: Admission control scheme for multimedia connections.

network, each class τ is represented by its effective bandwidth ϕ_τ. For admission control, we associate two *guard thresholds* $\Delta_{h,\tau}$ and $\Delta_{n,\tau}$ with each traffic class τ. A cell accepts an incoming handoff request of class τ if and only if the available bandwidth in that cell is greater than $\Delta_{h,\tau}$ plus the bandwidth of the connection. Otherwise, the handoff request is rejected and the connection is prematurely terminated. Similarly, a request for a new connection in a cell is accepted if and only if the available bandwidth in the cell is greater than $\Delta_{n,\tau}$ plus the bandwidth of the connection. Otherwise, the new connection request is rejected. Since premature termination of an ongoing connection is usually more undesirable than rejection of new connection request, $\Delta_{n,\tau} \geq \Delta_{h,\tau}$ for each traffic class τ.

The challenge is how to select the values of the guard thresholds such that most, if not all, handoff requests are accepted without significantly jeopardizing the probability of acceptance of a new request. In Section 3, we propose a strategy for selecting the values of the guard thresholds. Other strategies have been discussed in the literature. Before describing our strategy we briefly describe three different strategies from the literature. We refer to these strategies as Fixed, Static, and YL97. A comparison of the performance of our strategy relative to these strategies is given in Section 4.

2.1 Fixed(f) Strategy

In this strategy, each basestation sets aside $f\%$ of its capacity for dealing with handoff requests. This is achieved by choosing the guard threshold values to be $f\%$ of the cell's capacity. Specifically, if Γ_c is the capacity of cell c, then the basestation in c selects $\Delta_{h,\tau} = 0$ and $\Delta_{n,\tau} = f \cdot \Gamma_c$ for each traffic class τ.

2.2 Static(k) Strategy

The key limitation of the Fixed(f) strategy is that the threshold values are not directly based on the effective bandwidths of the connection requests. The Static(k) strategy, on the other hand, is cognizant of the effective bandwidths of the handoff requests.

In this strategy, the basestation is assumed to be aware of the steady fraction of connection requests for each traffic class τ. This fraction may be determined from historic traffic information available to the basestation. Let p_τ denote the fraction of connection requests for class τ. Then, the expected effective bandwidth for a handoff request is $\sum_{i=1}^{M} p_i \phi_i$. In Static($k$) strategy, each basestation selects $\Delta_{h,\tau} = 0$ and $\Delta_{n,\tau} = k \cdot \sum_{i=1}^{M} p_i \phi_i$ for each traffic class τ.

Note that, if all connection requests are identical, then this strategy is equivalent to selecting k guard channels.

2.3 YL97 Strategy

This strategy is based on the scheme presented in [1]. For comparison to the proposed strategy, this strategy has been modified slightly to deal with M classes of traffic. We first give an overview of the strategy proposed in [1] and then discuss our extension to deal with multiple traffic classes.

In this strategy, each basestation dynamically adapts the guard threshold values based on current estimates of the rate at which mobiles in the neighboring cells are likely to incur a handoff into this cell. The objective of the adaptation algorithm is to maintain a target block probability for handoff requests, despite temporal fluctuations in the connection request rate into the cell.

The determination of the guard threshold values is based on an analytic model which relates the guard threshold values to the blocking probabilities for handoff and new connection requests. This model requires the following key assumptions [1].

1. The arrival of new connection requests in a cell forms a Poisson process.

2. The arrival of handoff requests in a cell forms a Poisson process.

[1] These assumptions are not required for the strategy proposed in this paper.

```
Extended YL97 Strategy
    Let $N_{T,i} = \Gamma_c/\phi_1$.
    For $i = 1$ to $M$ do
        $N_{G,i} = f(N_{T,i}, \lambda_{n,i}, \lambda_{h,i}, \mu_i, B_h, B_n)$;
        $N_{T,i+1} = (N_{T,i} - N_{G,i})\phi_i$;
    Endfor
    $\Delta_{h,\tau} = 0$ for all $1 \leq \tau \leq M$
    $\Delta_{n,\tau} = \sum_{i=1}^{M} N_{G_i} \cdot \phi_i$ for all $1 \leq \tau \leq M$.
End.
```

Figure 2: Extended version of YL97 scheme to deal with multiple traffic classes.

3. The time spent by a connection in a cell is exponentially distributed.
4. The change in arrival rates is moderate in the sense that the network reaches steady state between any two changes in the arrival rate.

In this strategy, each basestation periodically queries neighboring basestations and computes an estimate of the rate at which handoff connection requests are expected to arrive in next update period. This estimate is derived from known stochastics of the connection duration times, cell residence times, and mobility patterns. The arrival of new connection requests is also estimated based on local measurements. Using expressions from queueing analysis, the basestation can then estimate the blocking probabilities for handoff and new connection requests as a function of the number of guard channels. From this function, the basestation computes the minimum number of guard channels required to meet the target blocking probabilities for handoff requests.

For comparison to the proposed strategy, we extend this strategy to deal with multiple traffic classes. To explain this extension, we need the following notations. Consider a typical cell c. Let $\lambda_{n,\tau}$ and $\lambda_{h,\tau}$ respectively be the estimated arrival rate of new and handoff connections of class τ in cell c for the next update period. Let μ_τ be the estimated departure rate of class τ connections in cell c. Also let, B_h and B_n denote the target blocking probabilities for handoff and new connection requests and let Γ_c be the total capacity of the cell. Furthermore, without loss of generality, assume that the effective bandwidths are such that $\phi_1 > \phi_2 > \cdots > \phi_M$. Figure 2 shows a pseudo-code of the extended version of the YL97 scheme. In this pseudo-code, the function $f()$ computes the minimum number of guard channels required to achieve the target handoff blocking probability exactly as in [1].

3 Dynamic `ExpectedMax` Strategy

Consider a typical cell c. Let t be the time of arrival of a new connection request in cell c. At time t, the basestation in cell c sends a query to the basestations in the neighboring cells requesting the information required to compute the guard threshold values. Once the guard threshold values are computed, the admission control scheme described in Figure 1 is used to determine whether or not to establish the new connection. Presented below is a formal description of the scheme used to compute the guard threshold values.

Ideally, the update of the guard threshold values in the proposed strategy must occur in a cell upon arrival of each new connection request. However, because of the associated communication and control overhead, it may not be possible in practice to update the threshold values so frequently. Therefore, in practice, basestations may update the guard threshold values once every $K \geq 1$ new connection requests, where K is a design parameter. Larger values of K means less overhead. However, since larger K means that the updates will be performed less frequently, the performance of the proposed strategy may worsen as compared to the ideal strategy. The effect of value of K on the performance of the proposed strategy is evaluated using a discrete-event simulator and the detailed results of this evaluation are shown in Section 4. The results basically show that impact on the performance is very small. Therefore, for ease of understanding, in the description of the proposed strategy, we assume that the update is performed upon arrival of each new connection request.

If accepted, let d be the expected duration of the new connection in cell c. Note that, the connection will leave cell c either due to completion or due to a handoff out of the cell. Therefore, the expected duration of the new connection in the cell can be estimated based on known stochastics of the unencumbered completion and cell residence times of connections. A technique for estimating the value of d is discussed later in this section. For now, assume that the value of d is known. Let m_τ be the number of class τ connections in cell c which are expected to either complete or incur a handoff out of the cell in the time interval $(t, t + d]$. Likewise, let n_τ be the expected number of connections of class τ in the neighboring cells which will incur a handoff into cell c in the time interval $(t, t + d]$. In practice, the values of m_τ and n_τ must be estimated by the cell and can therefore be inaccurate. However, for now, assume that their values are known exactly. After describing the basic idea of the proposed strategy, we describe a method for estimating the values of m_τ and n_τ.

Bandwidth is relinquished to cell c when a connection completes or incurs a handoff out of it. In contrast, additional bandwidth is required to accept an incoming handoff request. If sufficient bandwidth is not available for the incoming handoff request, the connection will have to be terminated prematurely. Similarly, additional bandwidth is required to accept a new connection in cell c. In this case, if sufficient bandwidth is not available, the connection is not established. This is usually preferable to prematurely terminating an established connection.

Define an *outgoing* τ-event (denoted by O) to be either a completion of a class τ connection or a handoff of a class τ connection from cell c. Similarly, define an *incoming* τ-event (denoted by I) to be a handoff of a class τ connection into cell c. Note that, a request for a new connection of class τ is not considered an *incoming* τ-event. This is because the threshold $\Delta_{n,\tau}$ for accepting a new connection request is set based on the expected bandwidth required to deal with the handoff requests; therefore, the computation of $\Delta_{n,\tau}$ depends only on handoffs and completions. Now consider the sequence of events which occur in cell c in the interval $(t, t+d]$. From the definition of m_τ and n_τ, we know that there will be $(m_\tau + n_\tau)$ events in this interval. Let $s \equiv a_1 a_2 \ldots a_{(m_\tau+n_\tau)}$ denote this sequence of events, where

$$a_l = \begin{cases} O & \text{if } l^{th} \text{ event is a completion or an outgoing handoff} \\ I & \text{if } l^{th} \text{ event is an incoming handoff.} \end{cases}$$

for $1 \leq l \leq (m_\tau + n_\tau)$. Furthermore, given s, let $X_{\tau,k}(s)$ denote the net change in the bandwidth allocated to class τ connections from time t to the end of k^{th} event in s. More formally, assuming $X_{\tau,0}(s) = 0$,

$$X_{\tau,k}(s) = \begin{cases} X_{\tau,k-1}(s) - \phi_\tau & \text{if } a_k \equiv O \\ X_{\tau,k-1}(s) + \phi_\tau & \text{otherwise.} \end{cases}$$

for $1 \leq k \leq (m_\tau + n_\tau)$. Define $Y_\tau(s) = \max\{X_{\tau,k}(s) : 0 \leq k \leq (m_\tau + n_\tau)\}$. Informally, $Y_\tau(s)$ is the maximum net change in the bandwidth allocated to class τ connections in $(t, t+d]$. Define $S(m_\tau, n_\tau)$ to be the set of all possible sequences of τ-events in $(t, t+d]$, i.e., $S(m_\tau, n_\tau)$ contains sequences of length $S(m_\tau + n_\tau)$ where each element in the sequence belongs to the set $\{I, O\}$ such that there are exactly n_τ I's in each sequence. More formally,

$$S(m_\tau, n_\tau) = \left\{a_1 \ldots a_{(m_\tau+n_\tau)} : |\{j : a_j = I\}| = n_\tau\right\}.$$

Since all sequences in $S(m_\tau, n_\tau)$ are equally likely to occur, the probability of occurrence of any particular sequence s is

$$p_\tau(s) = \frac{1}{|S(m_\tau, n_\tau)|}$$

and the expected value of $Y_\tau(s)$ is

$$\bar{Y}_\tau = \sum_{s \in S(m_\tau, n_\tau)} Y_\tau(s) p_\tau(s).$$

Intuitively, $\bar{Y}_\tau$ is the expected maximum net bandwidth that will be needed to deal with class τ handoff connections in the next update period. Therefore, $\sum_{\tau=1}^{M} \bar{Y}_\tau$ is expected maximum net bandwidth to deal with all handoff connections in the next update period. By setting $\Delta_{n,\tau} = \sum_{\tau=1}^{M} \bar{Y}_\tau$, this amount of bandwidth is effectively reserved for dealing with the incoming handoff requests in the interval $(t, t+d]$. This is the approach used in `ExpectedMax` strategy (see Figure 3). This idea is further clarified by the following example.

Example 1: Suppose at time t, $m_\tau = 2$ and $n_\tau = 3$. Then,

$$S(2,3) = \{IIIOO, IIOIO, IIOOI, IOIIO, IOIOI, OIIIO, OIIOI, IIOOI, IOOII, OOIII\}.$$

```
ExpectedMax
    $\Delta_{h,\tau} = 0$ for all $\tau$
    $\Delta_{n,\tau} = \sum_{j=1}^{M} \bar{Y}_j$ for all $\tau$
```

Figure 3: The proposed ExpectedMax strategy.

$s \in S(2,3)$	$Y_\tau(s)$	$p_\tau(s)$
IIIOO	$+3\phi_\tau$	0.1
IIOIO	$+2\phi_\tau$	0.1
IIOOI	$+2\phi_\tau$	0.1
IOIIO	$+2\phi_\tau$	0.1
IOIOI	$+\phi_\tau$	0.1
OIIIO	$+2\phi_\tau$	0.1
OIIOI	$+\phi_\tau$	0.1
IIOOI	$+2\phi_\tau$	0.1
IOOII	$+\phi_\tau$	0.1
OOIII	$+\phi_\tau$	0.1

Table 1: Values of $Y_\tau(s)$ for $s \in S(2,3)$.

Then,

$$\begin{aligned} X_{\tau,0}(IIOIO) &= 0 \\ X_{\tau,1}(IIOIO) &= \phi_\tau \\ X_{\tau,2}(IIOIO) &= 2\phi_\tau \\ X_{\tau,3}(IIOIO) &= \phi_\tau \\ X_{\tau,4}(IIOIO) &= 2\phi_\tau \\ X_{\tau,5}(IIOIO) &= \phi_\tau, \end{aligned}$$

and $Y_\tau(IIOIO) = 2\phi_\tau$. Intuitively, it means that if $Y_\tau(IIOIO)$ is reserved for dealing with incoming handoff requests and the actual sequence of events happens to be IIOIO, then all the incoming handoff requests can be accepted. However, IIOIO is only one out of the ten possible sequence of events and the bandwidth that will be required to accept all handoff requests will differ depending on the actual sequence of events. Table 1 shows the values of $Y_\tau(s)$ for all $s \in S(2,3)$ and the corresponding probability of occurrence of that s. From this table, it follows that the probability of the cell needing $+3\phi_\tau$ additional bandwidth to accept all incoming handoff requests is 0.1. Similarly, the probability of the cell needing $2\phi_\tau$ additional bandwidth is 0.5, and the probability of the cell needing ϕ_τ additional bandwidth is 0.4. If the basestation in the cell assumes that $3\phi_\tau$ bandwidth will be needed, then all incoming handoff requests can be accepted irrespective of the actual sequence. However, in most cases, it will be overallocating for handoff, because the probability of requiring $3\phi_\tau$ is only 0.1. Therefore, in the ExpectedMax strategy, the basestation assumes that will only need the expected value of $Y_\tau(s)$ instead of the maximum value of $Y_\tau(s)$. Note that, as a result, all incoming handoff requests are not guaranteed to be accepted. However, since $Y_\tau(s)$ is the maximum net bandwidth required if s is the actual sequence of events, reserving the expected value of $Y_\tau(s)$ will result in accepting most handoff requests. In this example, the basestation will assume that the bandwidth required to deal with handoffs is $3\phi_\tau \times 0.1 + 2\phi_\tau \times 0.5 + \phi_\tau \times 0.4 = 1.7\phi_\tau$. ■

3.1 Modification to `ExpectedMax` Strategy for Fairness

In the ExpectedMax strategy as described above, the handoff and new connection blocking probabilities will not be the same for the different classes of traffic. More specifically, classes with higher effective bandwidth will have higher handoff and new connection blocking probabilities as compared to those with smaller effective bandwidths. In some situations, it may be desirable to have comparable blocking probabilities irrespective of their effective bandwidths. To achieve fairness in blocking probabilities among all traffic classes, the guard threshold value for each class τ should be chosen as follows.

$$\begin{aligned}\Delta_{h,\tau} &= \phi_{\max} - \phi_\tau, \\ \Delta_{n,\tau} &= \phi_{\max} - \sum_{i=1}^{M} \bar{Y}_\tau,\end{aligned}$$

where $\phi_{\max} = \max_{1\leq i\leq M} \phi_i$.

The basic idea of this modification is to accept connection requests with smaller effective bandwidth if and only if the cell can accept a connection with the largest effective bandwidth. As a result, there will be an increase in the blocking probabilities for some traffic classes (especially, the ones with small effective bandwidth needs) and a decrease in the blocking probabilities of other traffic classes. The end result is that all classes will have comparable handoff blocking probabilities and comparable new connection blocking probabilities. However, since the guard threshold values for new connection requests includes the term $\sum_{i=1}^{M} \bar{Y}_\tau$, the blocking probabilities for handoff requests will be larger than that for new connection requests.

3.2 Computational issues in `ExpectedMax` strategy

3.2.1 Estimation of *d*

As described earlier, a connection leaves a cell either because it completes or because it incurs a handoff out of the cell. Therefore, the expected time spent by a connection in a cell can be derived from the probability distribution functions of the duration of class τ connection and the unencumbered cell residence times (i.e., residence time in a cell if the connection is of an infinite duration). Let $F_\tau(t)$ denote the probability distribution function of the duration of class τ connection. Also, let $R_\tau(t)$ denote the probability distribution function of the unencumbered cell residence time of a class τ connection. Let $r_\tau(t)$ denote the probability density function corresponding to $R_\tau(t)$.

Then, the probability distribution of the time spent by a connection in a cell is $1-(1-F_\tau(t))(1-R_\tau(t))$. The value of d can be estimated to be expected value of the time spent by a connection in a cell computed from this probability distribution function.

Note that, for the special case when the connection duration time is exponentially distributed and the unencumbered cell residence time is exponentially distributed, i..e,

$$\begin{aligned} F_\tau(t) &= 1 - e^{-\mu t} \\ R_\tau(t) &= 1 - e^{-\gamma t}, \end{aligned}$$

the probability distribution function of the time spent by a class τ connection in a cell is $1 - e^{-(\mu+\gamma)t}$. Therefore, the value of d can be estimated as $1.0/(\mu+\gamma)$.

3.2.2 Estimation of m_τ and n_τ

The conditional probability that a connection of class τ will incur a handoff in the interval $(t, t+d]$ given that it started at time u, entered the cell at time v, and is active at time t can be written as

$$\begin{aligned} H_\tau(u,v,t,d) &= \int_{w=t}^{t+d} \frac{\text{P[cell residence time} = w - v]}{\text{P[cell residence time} > t - v]} \cdot \frac{\text{P[conn. duration} > w - u]}{\text{P[conn. duration} > t - u]} dw \\ &= \int_{w=t}^{t+d} \frac{r_\tau(w-v)}{1 - R_\tau(t-v)} \cdot \frac{1 - F_\tau(w-u)}{1 - F_\tau(t-u)} dw, \end{aligned}$$

if $u \le v \le t$ and zero otherwise. Likewise, the conditional probability that a connection of class τ will neither incur a handoff in the interval $(t, t+d]$ nor complete in the interval $(t, t+d]$ given that it started at time u, entered the cell at time v, and is active at time t is

$$\begin{aligned} \overline{HC}_\tau(u,v,t,d) &= \frac{\text{P[conn. duration} > t+d-u, \text{cell res. time} > t+d-v]}{\text{P[conn. duration} > t-u, \text{cell res. time} > t-v]} \\ &= \frac{F_\tau(t+d-u)R_\tau(t+d-v)}{F_\tau(t-u)R_\tau(t-v)}, \end{aligned}$$

if $u \le v \le t$ and zero otherwise. Finally, the conditional probability that a connection of class τ will either complete or incur a handoff in the interval $(t, t+d]$ given that it started at time u, entered the cell at time v, and is active at time t is $1 - \overline{HC}_\tau(u,v,t,d)$.

Let G_τ^{c} be the set of all connections of class τ in cell c time t. Also, for each $c \in G_\tau^{\mathrm{c}}$, let u_c denote the time at which the connection started and v_c denote the connection c entered the cell under consideration. Note that, $v_c = u_c$ if the connected started in cell c. Then, in the `ExpectedMax` strategy, the basestation in cell c estimates m_T as

$$m_T = \left\lfloor \sum_{c \in G_T} (1 - \overline{HC}_T(u_c, v_c, t, d)) \right\rfloor ,$$

i.e., m_T is the expected number of connections to either complete or incur a handoff in time $(t, t+d]$.

The estimation of n_T requires interaction with neighboring cells. Let N_c denote the set of cells neighboring c. As explained earlier, the basestation in cell c sends a message to the basestation in each cell $j \in N_c$ requesting the information necessary to estimate n_T. The basestation then estimates

$$n_T = \left\lceil \sum_{j \in N_i} n_T^j \right\rceil ,$$

where n_T^j denotes the value returned from cell $j \in N_c$. The basestation in the neighboring cell j computes as follows. For each connection, $c \in G_T^j$, let $q_{j,c}(c)$ denote the conditional probability that the handoff will be to cell c given that connection c incurs a handoff. Then,

$$n_T^j = \sum_{c \in G_T^j} H(u_c, v_c, t, d) q_{j,c}(c).$$

The above expressions for estimating m_T and n_T hold for any given probability distribution function for connection duration time and cell residence time. For the special case, when the connection duration time is exponentially distributed and the cell residence time is also exponentially distributed, the above expressions become even simpler and are as shown below. These expressions are obtained by substitution and algebra in the above general expressions.

$$\begin{aligned}
F_T(t) &= 1 - e^{-\mu t} \\
R_T(t) &= 1 - e^{-\gamma t} \\
H_T(u, v, t, d) &= \frac{\gamma}{\gamma + \mu}(1 - e^{-(\gamma+\mu)d}) \\
\overline{HC}_T(u, v, t, d) &= e^{-(\gamma+\mu)d}.
\end{aligned}$$

3.2.3 Efficient computation of $\bar{Y}_T$

Recall that,

$$\bar{Y}_T = \sum_{s \in S(m_T, n_T)} Y_T(s) p_T(s).$$

If this expression is evaluated directly, the computational complexity is proportional to the cardinality of the set $S(m_T, n_T)$. We know that the cardinality of $S(m_T, n_T)$ is

$$|S(m_T, n_T)| = \frac{(m_T + n_T)!}{m_T! n_T!}$$

The main problem with this approach is that $|S(m_\tau, n_\tau)|$ can be large when m_τ and n_τ are large. We describe below a scheme to reduce the computational complexity of the above expression.

Recall that, $Y_\tau(s) = \max\{X_{\tau,k} : 0 \leq k \leq (m_\tau + n_\tau)\}$. Define $Z_\tau(s) = \max\{X_{\tau,k} : 1 \leq k \leq (m_\tau + n_\tau)\}$. Since $X_{\tau,0}(s) = 0$ for all s,

$$Y_\tau(s) = \max\{0, Z_\tau(s)\}.$$

Define $f(m_\tau, n_\tau, b)$ to be the number of sequences $s \in S(m_\tau, n_\tau)$ for which $Z_\tau(s) = b\phi_\tau$, That is,

$$f(m_\tau, n_\tau, b) = |\{s : s \in S(m_\tau, n_\tau), Z_\tau(s) = b\phi_\tau\}|$$

From the definition of $Z_\tau(s)$, the value of b can range from -1 to n_τ and

$$\bar{Y}_\tau = \sum_{b=-1}^{n_\tau} \frac{\max\{0, b\} \cdot \phi_\tau \cdot f(m_\tau, n_\tau, b)}{|S(m_\tau, n_\tau)|}. \tag{1}$$

Theorem 1: Given m_τ and n_τ such that $m_\tau + n_\tau \geq 1$, $f(m_\tau, n_\tau, b)$ is the solution of the following recursive equation.

$$f(m_\tau, 0, b) = \begin{cases} 1 & \text{if } m_\tau = 0, n_\tau = 1, b = 1 \\ 1 & \text{if } m_\tau > 0, n_\tau = 0, b = -1 \\ f(m_\tau - 1, n_\tau, 0) + f(m_\tau - 1, n_\tau, -1) & \text{if } m_\tau \geq 0, n_\tau > 0, b = -1 \\ f(m_\tau - 1, n_\tau, 1) & \text{if } m_\tau \geq 0, n_\tau > 0, b = 0 \\ f(m_\tau, n_\tau - 1, 0) + f(m_\tau, n_\tau - 1, -1) + f(m_\tau - 1, n_\tau, 2) & \text{if } m_\tau > 0, n_\tau > 0, b = 1 \\ f(m_\tau, n_\tau - 1, b - 1) + f(m_\tau - 1, n_\tau, b + 1) & \text{if } m_\tau > 0, n_\tau > 0, b > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Proof: The proof of this theorem is given in the appendix. ■

Let $\phi_{\min} = \min\{\phi_\tau : 1 \leq \tau \leq M\}$. Then, the maximum value of n_τ in cell c is $|N_c| \cdot \frac{\Gamma}{\phi_{\min}}$, where $|N_c|$ is the number of neighboring cells of c and Γ bandwidth of wireless link. Similarly, the maximum value of $m_\tau = \frac{\Gamma}{\phi_{\min}}$. Therefore, the values of $f(m_\tau, n_\tau, b)$ can be computed offline and stored in a table. The stored values can be used at runtime to compute $\bar{Y}_\tau$ using Equation 1.

4 Evaluation of ExpectedMax Strategy

In this section, we compare the performance of ExpectedMax strategy with that of other schemes in literature. The comparison is done using a C-based discrete-event wireless network simulator. The inputs to the simulator are a model of the wireless network and the characteristics/requirements of the multimedia traffic in this network. The outputs of the simulator include the blocking probabilities for handoff and new connection requests.

In this section, we compare the handoff and new connection blocking probabilities of four different strategies. The first strategy, labeled `Fixed`(5%), is a static scheme in which each basestation sets the threshold $\Delta_{n,\tau}$ to be 5% of its capacity for all τ. The second strategy, labeled `Static`(3) is also a static scheme in which each basestation sets the threshold $\Delta_{n,\tau}$ to be 3 times the average bandwidth requirement of the connection requests. This strategy requires knowledge of the relative occurrences of different traffic classes in the network. The third strategy is the extended version of YL97 scheme (see description in Section 2) with hard constraint. Finally, the fourth strategy is the proposed `ExpectedMax` strategy.

The performance of the four strategies is compared for three different type of networks. In all cases, the assumed topology of the wireless network is as shown in Figure 4. The other common aspects in all the three network types are as follows.

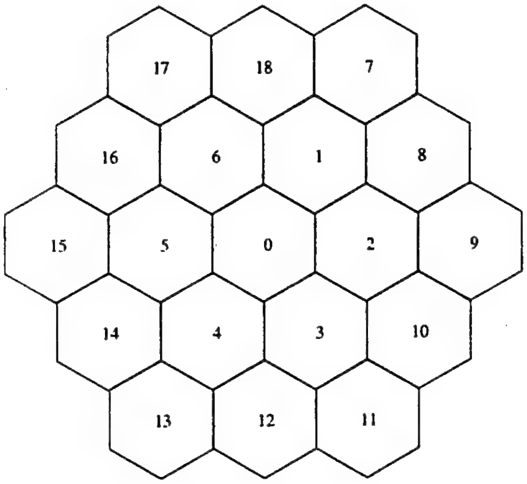


Figure 4: The simulated wireless network.

1. The arrival of new connection requests of class τ in each cell $\mathtt{c}$ is a Poisson process with rate $\lambda_{\tau,i}$. The rate $\lambda_{\tau,i}$ varies with time depending on the scenario.

2. The duration of each class τ connection request is selected from an exponential distribution with rate μ_τ. The duration of a connection is selected when it is first admitted into the network. Once determined, its value is fixed until the connection completes. The basestations, of course, do not use this information to make any decisions because in practice exact duration of a connection will not be known to the network.

3. The residence time of a class τ connection in a cell is chosen when the connection starts and when it incurs a handoff. Consider a connection which enters a cell at time v (i.e., it either started in the cell at time v or it incurred a successful handoff into the cell at time v). Let u' be the selected completion

Handoff Type	Probability
Suburb to suburb	0.25
Suburb to city	0.5
City to suburb	0.033
City to city	0.2
City to downtown	0.5
Downtown to city	0.166

Table 2: Steady-state handoff probabilities between cells in morning rush hour situation, for Network 1.

time for the connection. First, a random number w is generated from an exponential distribution with rate γ_T. If $v + w > u'$, then the connection completes in the cell at time u'. Otherwise, it incurs a handoff out of the cell at time $v + w$. Since, in practice, a basestation will not know the exact time of completion or handoff of a connection, this is assumed to be unknown to the basestation.

4. When a connection enters a cell (i.e., it either starts or it incurs a handoff into the cell), one of the neighboring cells is picked as a preferred cell for handoff. If the connection incurs a handoff (see discussion above), then handoff occurs to the preferred cell with 0.9 probability and with equal probability to one of the other neighboring cells. Since, in practice, a profile of a connection can be used to estimate the preferred handoff cell, the basestation is assumed to be aware of the preferred cell for each connection.

   The selection of the preferred cell for handoff is done as follows. As part of the input to the simulator, we specify parameters $q_{i,j}$ for each pair of adjacent cells i and j. $q_{i,j}$ represents the fraction of connections incurring a handoff from cell i which enter cell j. When a connection enters cell i, cell j is picked as a preferred cell for handoff with probability $q_{i,j}$.

4.1 Network 1

In this network type, we simulate the wireless network of Figure 4 with cell 0 in a downtown region, cells 1–6 in the city and cells 7-18 in the suburbs. We first consider the morning rush hour scenario in which most users are moving towards the downtown area from the suburbs and the city by selecting the parameters as shown in Table 2.

There are three classes of traffic in this network. We refer to them as class 0, class 1 and class 2. The parameters for class 2 connections are similar to that of a typical cellular phone conversation. In particular, the bandwidth requirement of a class 2 connection is 64 Kbps and its mean duration is assumed to be 150 seconds. Class 0 and Class 1 require much higher bandwidths and they also last longer on the average. This is because, in practice, users of higher bandwidth connections like video conferencing are typically connected for much longer duration as compared to typical voice connection. Specifically, the bandwidth requirements of class 0 and class 1 connections are respectively eight and four times that of a class 2 connection. Moreover, the mean duration of class 0 and class 1 connections are respectively 25 and 5 times that of a class 2 connection.

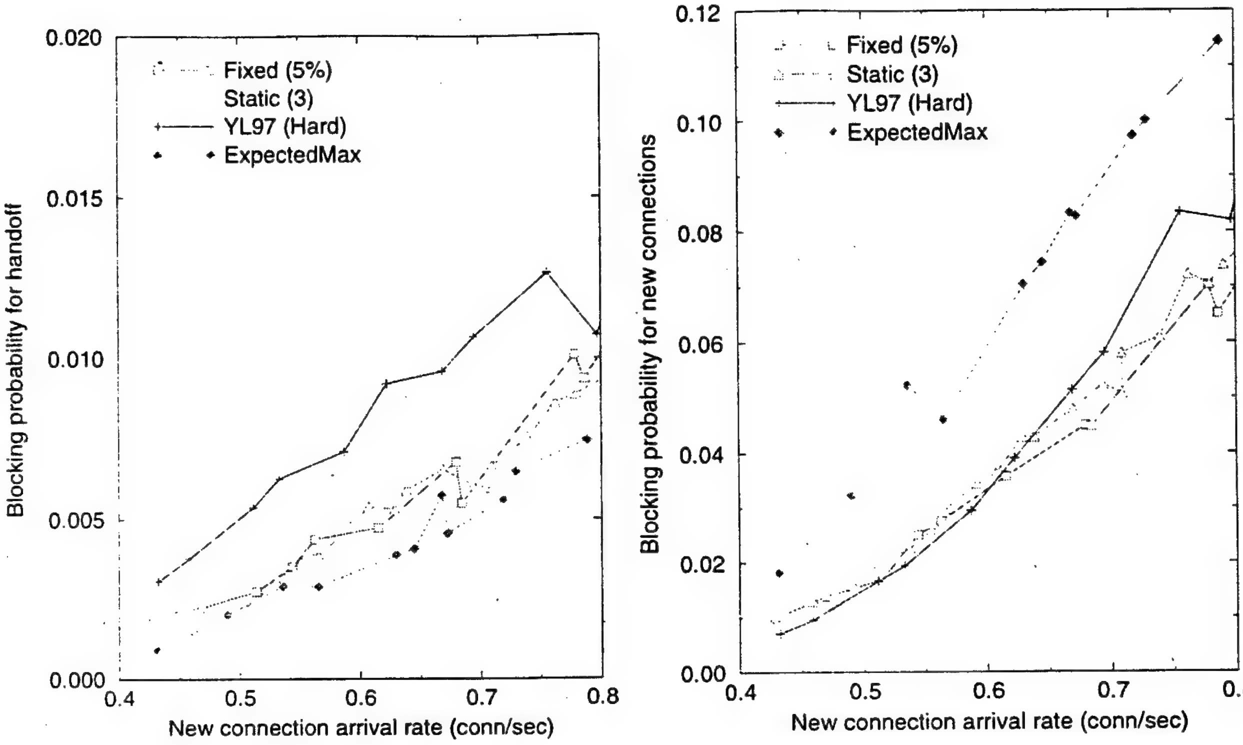
In each cell, the arrival rate of each class of connection increases in the first half of our simulation and decreases in the second half. This corresponds to a typical increase in the call arrival rate from say 6:00 a.m. to 8:00 a.m. and then a decrease in the call arrival rate from 8:00 a.m. to 10:00 a.m. The increases in the call arrival rate in the various cells do not occur at the same rate. In the first half of the simulation, approximately once every 24 minutes, the call arrival rate in a cell is increased by a random factor chosen from an uniform distribution between 1.0 and 1.4. Similarly, in the second half of the simulation, approximately once every 24 minutes, the call arrival rate in a cell is decreased by a random factor chosen from an uniform distribution between 1.0 and 1.4.

There is usually a difference in the new connection request arrival rate between a downtown cell and a suburb cell. To account for this, the new connection request arrival rate in downtown is assumed to be 40% higher on the average than in the suburb. Likewise, new connection request arrival rate in the city is on the average 20% higher than in the suburb. Similarly, in practice, there is also likely to be difference between the cell residence times of a connection for downtown, city, and suburb. For instance, cellular phone users in a downtown are more likely to remain in downtown as compared to city or suburb. To account for this the mean unencumbered cell residence times of each connection in downtown and city are respectively assumed to be 100% and 33% longer than that in the suburb.

Furthermore, since the cells differ considerably in the arrival rate of handoff and new connection requests, we assume that the total bandwidth available in the cells differ correspondingly. Specifically, we assume that the total bandwidth available in downtown is twice that of that of the suburb and 25% more than that of the city. Furthermore, in downtown, we assume that the total bandwidth available is adequate to simultaneously support at most twenty class 0 connections. The difference in the cell capacity can be achieved in practice by allocating more channels to the downtown cell as compared to the city and suburb. For example, 20 different frequencies can be assigned to the downtown cell, 16 frequencies to the city cells, and 10 frequencies to the suburb cells to achieve the above variation in cell capacity.

Figure 5(a) and 5(b) respectively show the blocking probabilities for handoff and new connection requests as a function of mean new connection request arrival rate for the four different strategies. Since this arrival rate increases in the first half and decreases in the second half of the simulation, the mean arrival rate is computed as the total number of connections which arrive during the simulation divided by the duration of the simulation.

First observe that, as expected, the blocking probabilities for handoff and new connection request increase in all strategies with increase in the arrival rate of new connection request. The performance of `Fixed(5%)` and `Static(3)` schemes are better than that of YL97 strategy. Importantly, note that the blocking probabilities for handoff request is the smallest for the proposed `ExpectedMax` strategy. The maximum load for which the target handoff blocking probability of 0.01 is achieved, is largest for our `ExpectedMax` strategy.



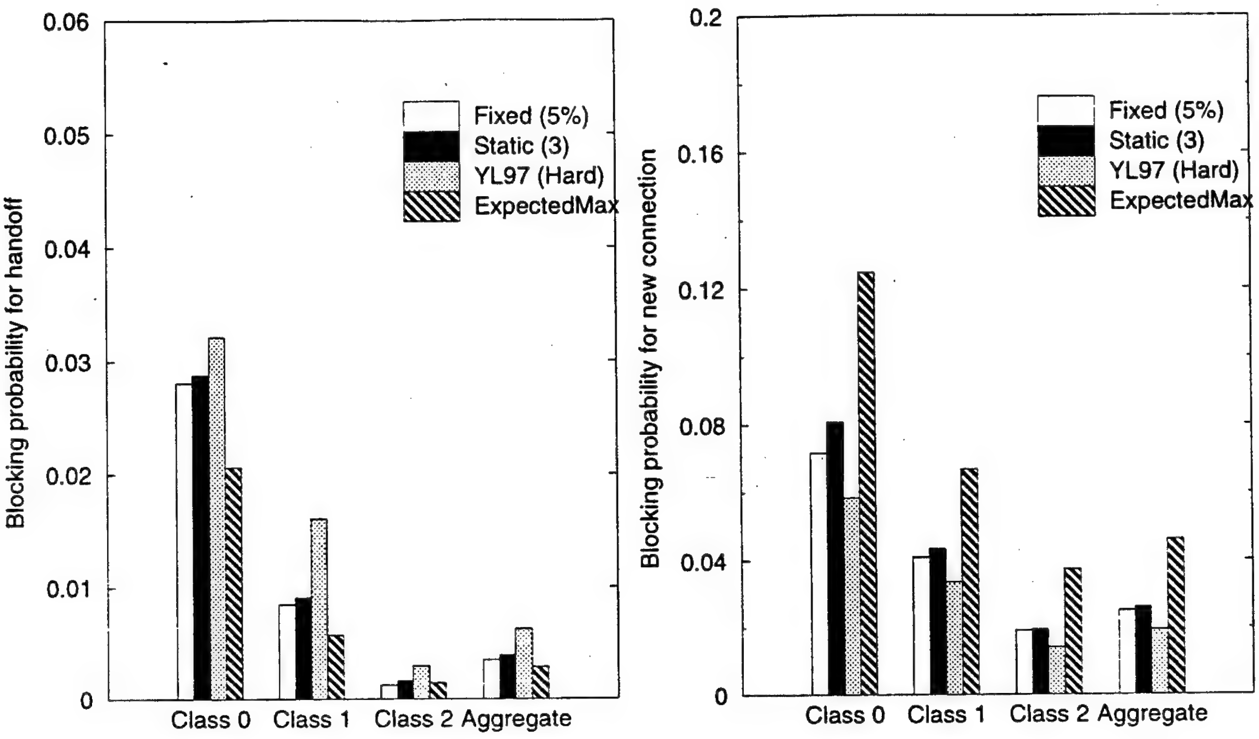
(a) Handoff

(b) New connection

Figure 5: Blocking probabilities in different strategies as a function of load in Network 1 with morning rush hour traffic.

Figure 5(b) shows that the new connection blocking probabilities are lower for all schemes compared to `ExpectedMax`. As mentioned earlier, there is an obvious tradeoff in blocking probabilities for handoff versus new connections. The network designer needs to make a decision regarding the amount of penalty he is willing to accept in terms of higher new connection blocking probability. For instance, the designer can decide on target probabilities of 0.005 and 0.10 respectively for handoff and new connections respectively. Our scheme targets such a situation and strives to closely estimate handoff resource requests to actual future requests.

Figures 6 and 7 show the blocking probabilities for the three individual classes, for two different system loads (medium and high). The medium load corresponds to a new connection arrival rate of 0.53 connections/sec whereas the high load corresponds to an arrival rate of 0.672 connections/sec. The `ExpectedMax` strategy consistently results in lower handoff blocking probabilities for all the three classes in both cases. Note that, class 2 has typically very low blocking probability, while class 0 has higher blocking probability. This is expected since class 2 has lower bandwidth requirements. This is also attractive to network service providers where voice (class 2) typically is the mainstream application compared to video (class 0).



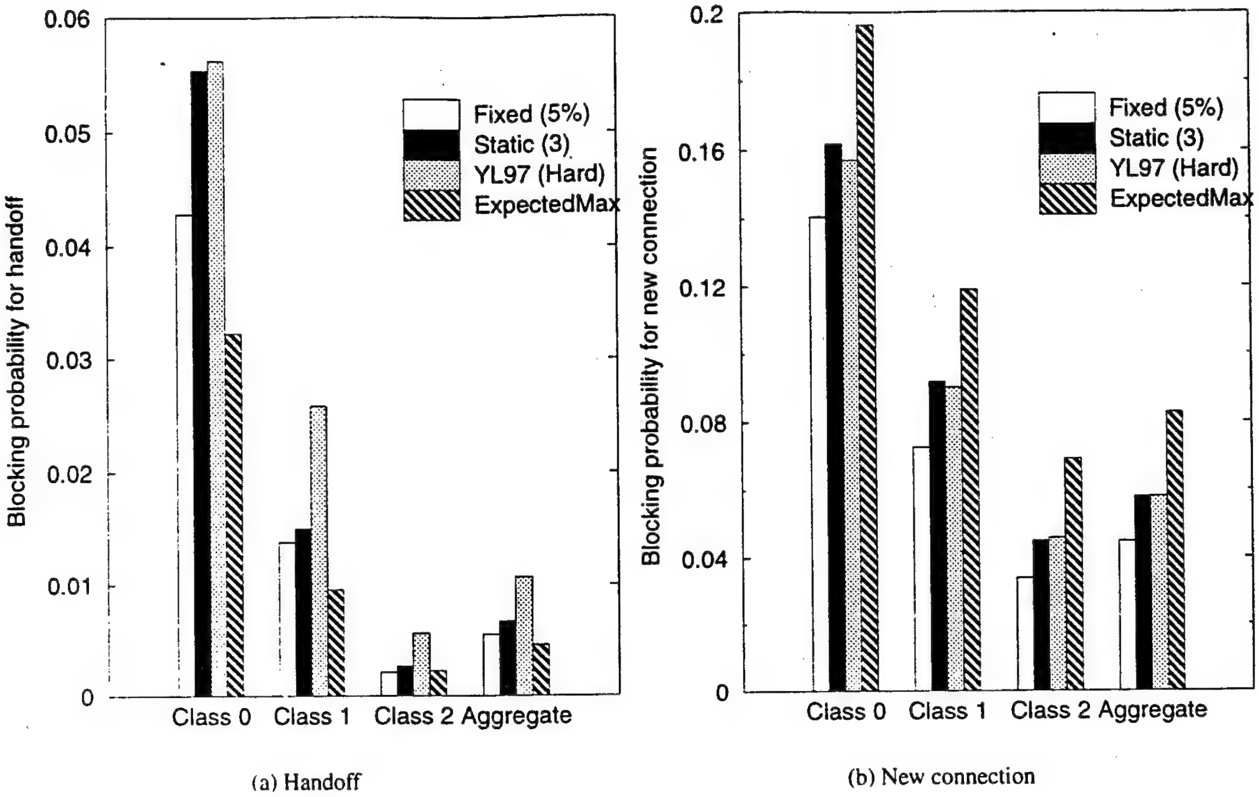
(a) Handoff

(b) New connection

Figure 6: Blocking probabilities for the traffic classes under medium load in Network 1 with morning rush hour traffic.

Update Frequency: As described earlier, ideally each basestation must get updated information about expected handoffs from neighboring upon arrival of each new connection request. The overhead of such frequent updates is clearly very high. Our objective is to minimize the frequency of update while achieving accurate estimates of handoff resource requests. Figure 8 shows the variation in blocking probabilities as update rate is decreased at one particular load, namely 0.672 conn/sec new connection arrival rate. The x-axis in Figure 8 shows the ratio of the mean new connection inter-arrival rate and the mean inter-update rate. A ratio of 1 means that, on the average, updated information is obtained from neighboring nodes for each incoming new connection request (i.e., approximately the ideal update rate). A ratio of 100 means that, the average, updated information is obtained from neighboring nodes once every 100 new connection requests. Observe that, the blocking probabilities are fairly steady even when the update rate is reduced to approximately 1/100 of the ideal update rate. This indicates that the proposed `ExpectedMax` strategy can be implemented without much overhead in terms of frequent updates between basestations.

Evening rush hour: To provide more perspective, we also simulated the condition where users are moving away from downtown towards city and suburbs. Specifically, the mobility parameters for are as shown in Table 3.



(a) Handoff

(b) New connection

Figure 7: Blocking probabilities for the traffic classes under high load in Network 1 with morning rush hour traffic.

Here, we assume that, on the average, arrival rate of new connection requests in downtown is 2.5 times that in suburb, and in the city in 2.0 times that in the suburb. As before, the arrival rates in each cell for traffic type increases in the first half of the simulation and decreases in the second half. The relative behavior of all strategies are very similar to that discussed for morning rush hour situation in Network 1 (see Figure 9). In particular, the proposed `ExpectedMax` strategy has the least handoff blocking probability. Correspondingly, it has the highest new connection request blocking probability. As stated earlier, since premature termination of established connection requests is probably more undesirable than rejection of new connection request, the proposed `ExpectedMax` strategy seems to be better even in this situation.

4.2 Network 2

The wireless network and the mobility pattern in this network are exactly as in Network 1. However, instead of three different classes of traffic, all connections requests are of the same class in this network. Specifically, the parameters of all the connection requests are that of a typical cellular phone conversation (i.e., class 2 connection of 1). This network was used in the evaluation described in [1] and is included here for comparison.

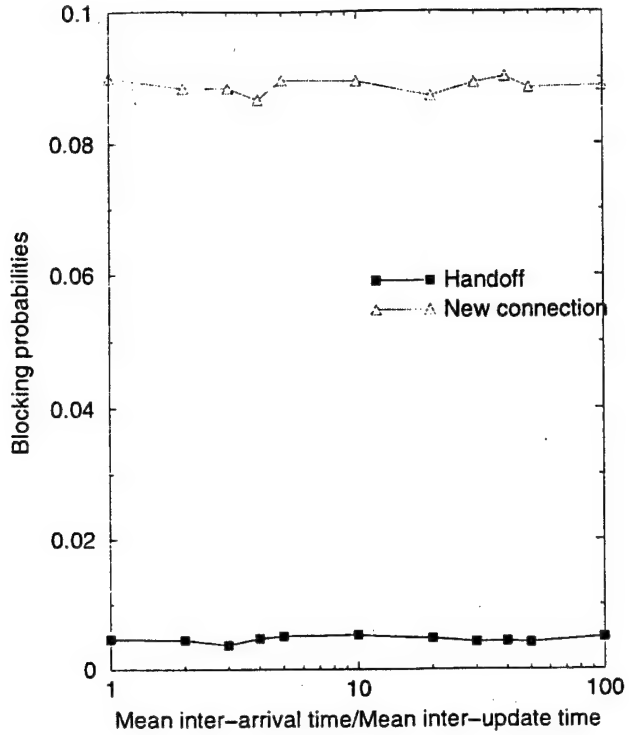


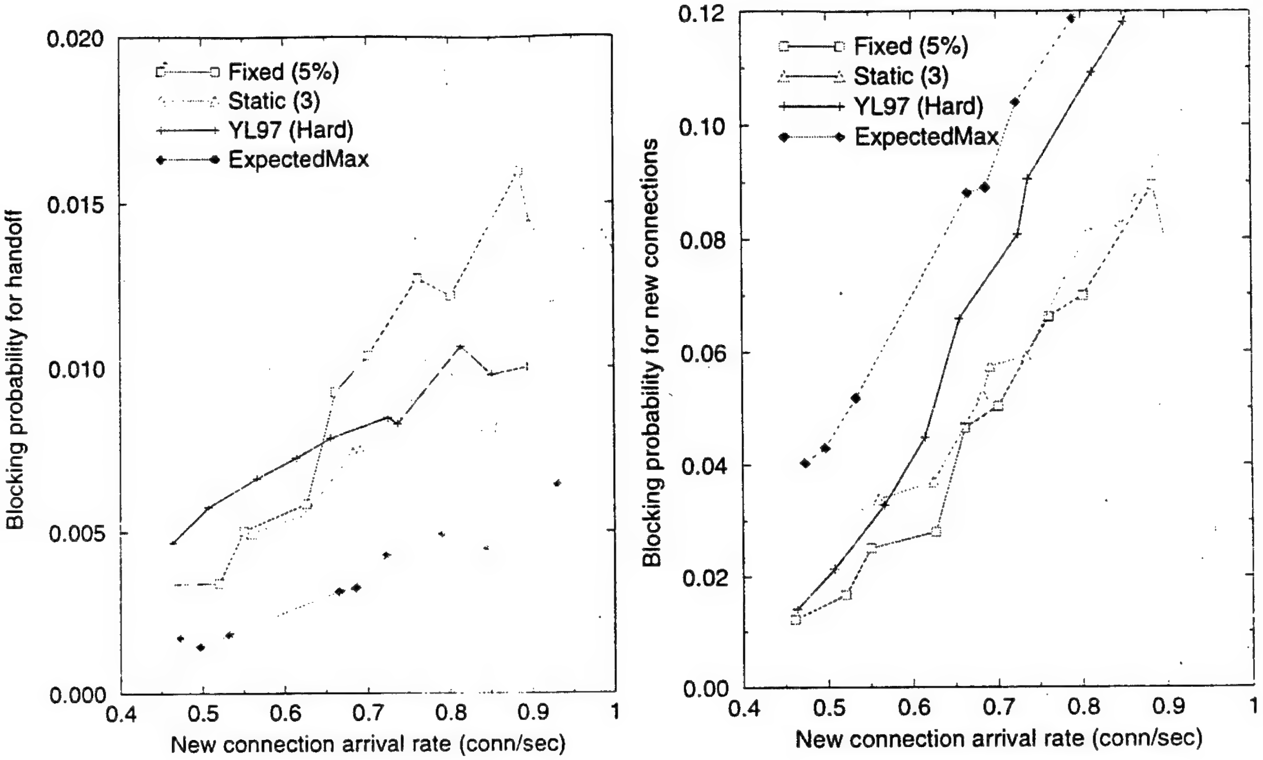
Figure 8: Blocking probabilities as update is reduced in Network 1 at high load.

As in Network 1, the arrival rate of new connection requests increases in the first half of the simulation and then decreases in the second half. The method used to increase and the decrease the rate of arrival of connections is exactly as in Network 1. Also, as in Network 1, the average call arrival rate in a downtown is 40% higher than that in suburb. The call arrival rate in the city is on the average 20% higher than in the suburb. Moreover, to account for the differences in the residence times among cells, mean unencumbered cell residence time in downtown (city) is assumed to be twice (1.33 times) that in suburb.

Figure 10 shows the variation in the blocking probabilities in the different strategies when arrival rate of new connection requests is increased. Unlike in Network 1, the `Static`(3) and the `Fixed`(5%) strategies

Handoff Type	Probability
Suburb to suburb	0.45
Suburb to city	0.05
City to suburb	0.25
City to city	0.1
City to downtown	0.05
Downtown to city	0.166

Table 3: Steady state handoff probabilities between cells in evening rush hour situation, for Network 1.

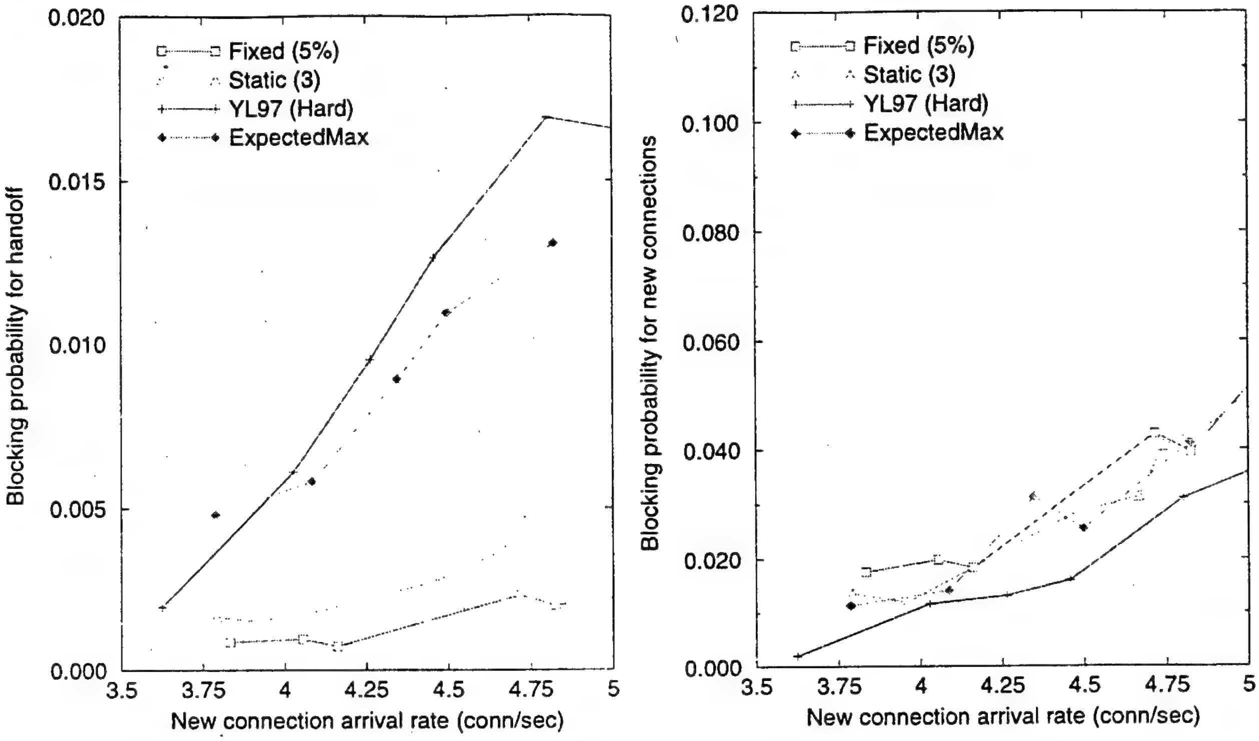


(a) Handoff

(b) New connection

Figure 9: Blocking probabilities in different strategies as a function of load in Network 1 with evening rush hour traffic.

perform quite well in this network. They have the smallest handoff blocking probability as compared to the dynamic schemes. The main reason is that the average bandwidth per connection is very small (approximately 0.0064) and therefore reserving 5% results in over-reserving resources for handoffs. Since the `Static`(3) reserves for at most three handoff connections, its blocking probability is higher when compared to `Fixed`(5%). The handoff blocking probabilities in the proposed `ExpectedMax` strategy are smaller than that of the YL97 schemes. However, the difference between the two schemes seems to be much less in this network as compared to in Network 1. The new connection blocking probabilities for the YL97 schemes are much better than that for `ExpectedMax` strategy. The results here seem to indicate that with a low-bandwidth homogeneous traffic network, a static scheme will be sufficient to obtain good performance. As mentioned earlier, static schemes do not require the overhead associated with basestation updates. As shown in the previous section, there is significant advantage in multiple class networks, with diverse traffic requirements.



(a) Handoff

(b) New connection

Figure 10: Blocking probabilities in different strategies as a function of load in Network 2.

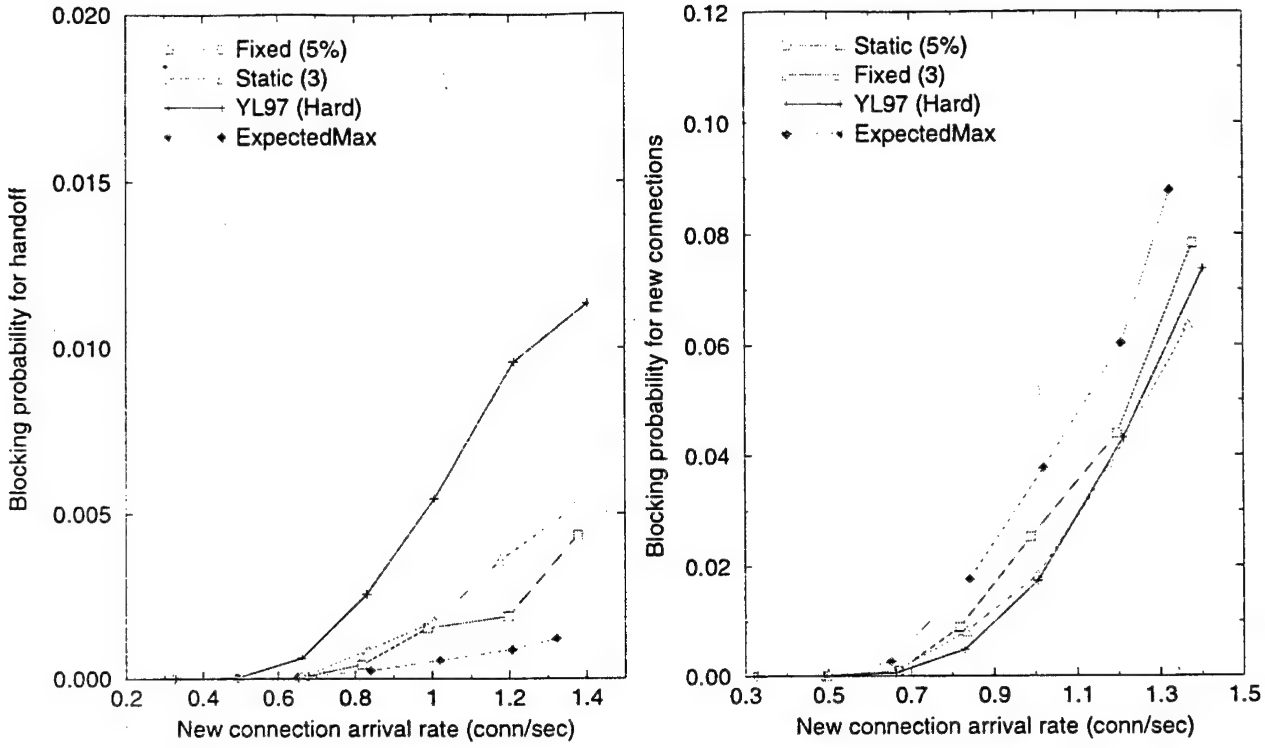
4.3 Network 3

This network is meant to capture a uniform network where all cells are identical in terms of traffic flow and the probabilities of moving between cells is uniform. That is, a mobile is equally likely to move to any of the neighboring cells. The parameters of the traffic classes are as in Network 1.

Here again, Figure 11 shows the variation in handoff and new connection blocking probabilities for different loads. As in Networks 1 and 2, the handoff blocking probability is the least for the `ExpectedMax` strategy. In this network, YL97 strategies are worse than the `Fixed`(5%) strategies and the `Static`(3) schemes for both handoff and new connection blocking probabilities. This shows that our `ExpectedMax` strategy is well adapted to different network conditions and traffic patterns.

5 Conclusion

This paper addressed the problem of providing resources to mobile connections during handoff between basestations. The network is assumed to be cell-based with support for diverse traffic types. The goal is to



(a) Handoff

(b) New connection

Figure 11: Blocking probabilities in different strategies as a function of load in Network 3.

estimate the requirements for resources from mobiles that are currently in neighboring cells and that might potentially move to the current cell. This estimate is then used to appropriately reserve resources in the current cell for potential handoff connections. A dynamic strategy that uses the estimated holding times and mobility pattern information is proposed in the paper. The performance of this strategy is compared to two fixed or static schemes that do not use any dynamic information, and to a scheme proposed earlier. The results were studied for three different type of networks. The performance metrics studied are blocking probabilities for handoff and for new connections. A fundamental assumption is that the network designer desires a lower handoff blocking probability. This requirement might result in higher blocking probability for new connections. A suitable tradeoff is required based on the network service objectives.

The results show that our proposed ExpectedMax strategy consistently achieves lower handoff blocking probability than all other schemes. The results also show that the dynamic estimation can be achieved without significant overhead in terms of control communication between basestations. Further work in this area includes translating the high-level resource allocations into scheduling at the multiple access level using an access protocol such as described in [16] to ensure quality-of-service.

A Proof of Theorem 1

Proof of Theorem 1: The theorem is proved by considering the individual cases in Equation 2, namely,

- Case 1: $m_\tau = 0, n_\tau = 1, b = 1$
- Case 2: $m_\tau > 0, n_\tau = 0, b = -1$
- Case 3: $m_\tau \geq 0, n_\tau > 0, b = -1$
- Case 4: $m_\tau \geq 0, n_\tau > 0, b = 0$
- Case 5: $m_\tau > 0, n_\tau > 0, b = 1$
- Case 6: $m_\tau > 0, n_\tau > 0, b > 1$

Case 6 is the common case. Cases 1–5 are in some sense boundary cases. Therefore, we prove Case 6 in detail and outline the key proof steps for other cases. Some of the cases need the following definitions and observation.

Let $S_I(m_\tau . n_\tau) \subseteq S(m_\tau, n_\tau) = \{s : s \in S(m_\tau, n_\tau), s \equiv Is'\}$ be the set of all sequences in $S(m_\tau, n_\tau)$ which start with I. Likewise, let $S_O(m_\tau, n_\tau) \subseteq S(m_\tau, n_\tau) = \{s : s \in S(m_\tau, n_\tau), s \equiv Os'\}$ be the set of all sequences in $S(m_\tau . n_\tau)$ which start with O. Let $f_I(m_\tau, n_\tau, b)$ be the number of sequences $s \in S_I(m_\tau, n_\tau)$ such that $Z_\tau(s) = b\phi_\tau$. Similarly, let $f_O(m_\tau, n_\tau, b)$ be the number of sequences $s \in S_O(m_\tau, n_\tau)$ such that $Z_\tau(s) = b\phi_\tau$. Observe that

$$f(m_\tau, n_\tau, b) = f_I(m_\tau, n_\tau, b) + f_O(m_\tau, n_\tau, b). \tag{3}$$

Case 6: $m_\tau > 0. n_\tau > 0, b > 1$

By definition any $s \in S_I(m_\tau, n_\tau)$, can be written as Is' for some s'. Thus, for $s \in S_I(m_\tau, n_\tau)$, $X_{\tau,1} = \phi_\tau$ and

$$\begin{aligned} Z_\tau(s) &= \max\{0, X_{\tau,1}, \max_{2 \leq k \leq |s|} X_{\tau,k}\} \\ &= \max\{\phi_\tau, \max_{2 \leq k \leq |s|} X_{\tau,k}(s)\} \\ &= \max\{\phi_\tau, \phi_\tau + \max_{1 \leq k \leq |s'|} X'_{\tau,k}(s')\} \\ &= \max\{\phi_\tau, \phi_\tau + Z_\tau(s')\}. \end{aligned}$$

Therefore, for $b > 1$,

$$Z_\tau(s) = b \cdot \phi_\tau \quad \text{iff} \quad Z_\tau(s') = (b-1)\phi_\tau.$$

Hence, $f_I(m_\tau, n_\tau, b) = f_I(m_\tau, n_\tau - 1, b - 1)$.

Similarly, by definition any $s \in S_O(m_\tau, n_\tau)$, can be written as Os' for some s'. Thus, for $s \in S_O(m_\tau, n_\tau)$, $X_{\tau,1} = -\phi_\tau$ and

$$\begin{aligned} Z_\tau(s) &= \max\{0, X_{\tau,1}, \max_{2\leq k\leq |s|} X_{\tau,k}\} \\ &= \max\{0, \max_{2\leq k\leq |s|} X_{\tau,k}(s)\} \\ &= \max\{0, -\phi_\tau + \max_{1\leq k\leq |s'|} X'_{\tau,k}(s')\} \\ &= \max\{0, -\phi_\tau + Z_\tau(s')\}. \end{aligned}$$

Therefore, for $b > 1$,

$$Z_\tau(s) = b \cdot \phi_\tau \quad \text{iff} \quad Z_\tau(s') = (b+1)\phi_\tau.$$

Hence, $f_O(m_\tau, n_\tau, b) = f_O(m_\tau - 1, n_\tau, b+1)$. The theorem then follows for this case from Equation 3.

Case 1: $m_\tau = 0, n_\tau = 1, b = 1$

Since $m_\tau = 0$ and $n_\tau = 1$, In this case, the set $S(m_\tau, n_\tau)$ contains only one element, namely the sequence I. Therefore, $Z_\tau(s) = \phi_\tau$ for all $s \in S(m_\tau, n_\tau)$. That is, $f(0,1,1) = 1$ and $f(0,1,b) = 0$ for all $b \neq 1$.

Case 2: $m_\tau > 0, n_\tau = 0, b = -1$

In this case, the set $S(m_\tau, 0)$ contains only one sequence, namely a sequence of m_τ O's. Therefore, $Z_\tau(s) = X_{\tau,1}(s) = -\phi_\tau$ for all $s \in S(m_\tau, 0)$. That is, $f(m_\tau, 0, -1) = 1$ and $f(m_\tau, 0, b) = 0$ for all $b \neq -1$ and all $m_\tau > 0$.

Case 3: $m_\tau \geq 0, n_\tau > 0, b = -1$

For any $s \in S_I(m_\tau, n_\tau)$, $X_{\tau,1}(s) = \phi_\tau$. Therefore, $Z_\tau(s) \geq \phi_\tau$ for all for all $s \in S_I(m_\tau, n_\tau)$ and $f_I(m_\tau, n_\tau, -1) = 0$.

By definition $s \in S_O(m_\tau, n_\tau)$, can be written as Os' for some s' and

$$Z_\tau(s) = -\phi_\tau \quad \text{iff} \quad Z_\tau(s') = 0 \ \text{ or } \ Z_\tau(s') = -\phi_\tau.$$

Therefore, $f_O(m_\tau, n_\tau, -1) = f_O(m_\tau - 1, n_\tau, 0) + f_O(m_\tau - 1, n_\tau, -1)$. The theorem follows for this case from Equation 3.

Case 4: $m_\tau \geq 0, n_\tau > 0, b = 0$

For any $s \in S_I(m_\tau, n_\tau)$, $X_{\tau,1} = \phi_\tau$. Therefore, $Z_\tau(s) \geq \phi_\tau$ for all $s \in S_I(m_\tau, n_\tau)$ and $f_I(m_\tau, n_\tau, 0) = 0$.

Similarly, any $s \in S_O(m_\tau, n_\tau)$, can be written as Os' for some s'. Thus, for $s \in S_O(m_\tau, n_\tau)$,

$$Z_\tau(s) = 0 \quad \text{iff} \quad Z_\tau(s') = \phi_\tau,$$

and therefore $f_O(m_\tau, n_\tau, 0) = f_O(m_\tau - 1, n_\tau, 1)$. The theorem follows for this case from Equation 3.

Case 5: $m_T > 0, n_T > 0, b = 1$
By definition, any $s \in S_I(m_T, n_T)$, can be written as Is' for some s'. Thus, for $s \in S_I(m_T, n_T)$,

$$Z_T(s) = \phi_T \quad \text{iff} \quad Z_T(s') = 0 \ \text{ or } \ Z_T(s') = -\phi_T.$$

Therefore, $f_I(m_T, n_T, 1) = f_I(m_T, n_T - 1, 0) + f_I(m_T, n_T - 1, -1)$.

Likewise, any $s \in S_O(m_T, n_T)$, can be written as Os' for some s' and therefore

$$Z_T(s) = \phi_T \quad \text{iff} \quad Z_T(s') = 2\phi_T.$$

Hence, $f_O(m_T - 1, n_T, 1) = f_O(m_T - 1, n_T, 2)$. The theorem follows for this case from Equation 3.

The theorem follows from Cases 1–6. ■

References

[1] O. T. W. Yu and V. C. M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1208–1225, Sept. 1997.

[2] K. Pahlavan and A. H. Levesque, "Wireless data communications," *Proceedings of the IEEE*, vol. 82, pp. 1398–1430, Sept. 1994.

[3] M. Naghshineh (Guest Ed.), "Wireless ATM: Special Issue." IEEE Personal Communications, Aug. 1996.

[4] T. Hsing, D. C. Cox, L. F. Chang, and T. Van Landegren (Guest Ed.), "Wireless ATM: Special Issue." IEEE Journal on Selected Areas in Communications, Jan. 1997.

[5] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. Srivastava, and J. A. Trotter, "SWAN: A mobile multimedia wireless network," *IEEE Personal Communications*, pp. 18–33, Apr. 1996.

[6] P. Agrawal, D. K. Anvekar, and B. Narendran, "Channel management policies for handovers in cellular networks," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 97–110, 1996.

[7] J. Daigle and N. Jain, "A queueing system with two arrival streams and reserved servers with application to cellular telephone," in *Proceedings of INFOCOM*, pp. 2161–2167, Apr. 1992.

[8] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systesm with prioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. 3, pp. 77–91, Aug. 1986.

[9] M. Naghshineh and M. Schwartz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, pp. 711–717, May 1996.

[10] C. H. Yoon and C. K. Un, "Performance of personal portable radio telephone systems with and without guard channels," *IEEE Journal on Selected Areas in Communications*, vol. 11, pp. 911–917, Aug. 1993.

[11] M. M. Zonoozi and P. Dassanayake, "User mobility modeling and characterization of mobility patterns," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1239–1252, Sept. 1997.

[12] A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admisssion control of high-speed networks," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 329–343, June 1993.

[13] R. Guerin, H. Ahmadi, and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, pp. 968–981, Sept. 1991.

[14] F. P. Kelly, "Effective bandwidths at multi-type queues," *Queueing systems*, vol. 9, pp. 5–15, 1991.

[15] J.-Y. L. Boudec, "Network calculus made easy," Tech. Rep. EPFL-DI 96/218, http://lrcwww.epfl.ch, Dec. 1996.

[16] K. M. Sivalingam, M. B. Srivastava, P. Agrawal, and J.-C. Chen, "Low-power Access Protocols Based on Scheduling for Wireless and Mobile ATM Networks," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, pp. 429–433, Oct. 1997.

UNDER REVIEW, ACM WINET JOURNAL

Performance Comparison of Battery Power Consumption in Wireless Multiple Access Protocols

Jyh-Cheng Chen[1], Krishna M. Sivalingam[2], Prathima Agrawal[3], and Shalinee Kishore[4] *
[1]Department of Electrical & Computer Engineering, State University of New York at Buffalo, Buffalo, NY 14260
[2]School of Electrical Engineering & Computer Science, Washington State University, Pullman, WA 99164
[3]Networked Computing Technology Department, AT&T Labs, Whippany, NJ 07981
[4]WINLAB, Rutgers University, Piscataway, NJ 08855

Abstract

Energy efficiency is an important issue in mobile wireless networks since the battery life of mobile terminals is limited. Conservation of battery power has been addressed using many techniques such as variable speed CPUs, flash memory, disk spindowns, and so on. We believe that energy conservation should be an important factor in the design of networking protocols for mobile wireless networks. In particular, this paper addresses energy efficiency in medium access control (MAC) protocols for wireless networks. The paper develops a framework to study the energy consumption of a MAC protocol from the transceiver usage perspective. This framework is then applied to compare the performance of a set of protocols that includes IEEE 802.11, EC-MAC, PRMA, MDR-TDMA, and DQRUMA†. The performance metrics considered are transmitter and receiver usage times for packet transmission and reception. The time estimates are then combined with power ratings for a Proxim RangeLAN2 radio card to obtain an estimate of the energy consumed for MAC related activities. The analysis here shows that protocols that aim to reduce the number of contentions perform better from an energy consumption perspective. The receiver usage time, however, tends to be higher for protocols that require the mobile to sense the medium before attempting transmission. The paper also provides a set of principles that could be applied when designing access protocols for wireless networks.

Keywords: wireless networks, medium access control (MAC), multiple access protocols, energy efficiency, low-power operation.

1 Introduction

This paper addresses the issue of energy conservation in medium access control (MAC) protocols for wireless multimedia networks. Third generation wireless networks will be expected to carry diverse multimedia traffic types. A number of access protocols have been proposed to support multimedia traffic [1–6]. These

*A preliminary version of this paper was presented at IEEE INFOCOM 1998 in San Francisco, CA. CONTACT: Krishna Sivalingam, School of Electrical Engg. and Computer Science, Washington State University, Pullman, WA 99164. Phone: 509 335 6602, Fax: 509 335 3818. Email: krishna@eecs.wsu.edu. The authors can be reached at jcchen@eng.buffalo.edu, krishna@eecs.wsu.edu, pa@research.att.com, and kishore@ece.rutgers.edu. Part of the research was supported by Air Force Office of Scientific Research grant F-49620-97-1-0471.

†EC-MAC: energy-conserving MAC. PRMA: packet reservation multiple access. MDR-TDMA: multiservices dynamic reservation TDMA. DQRUMA: distributed-queuing request update multiple access.

protocols typically address network performance metrics such as throughput, efficiency, and packet delay. The comparison for some of the wireless MAC protocols based on the perspective of quality-of-services (QoS) can be found in [7]. We believe that energy consumption at the MAC level should also be an important consideration in the design of the MAC protocol for mobile wireless networks. The premise is that mobiles will always have limited power, whereas the wired base stations will have virtually unlimited power. In order to understand the design issues, we provide in this paper a comparison of the energy-efficiency of a set of MAC protocols that includes the IEEE 802.11 standard [8].

The paper considers an infrastructure network where a base station coordinates access to one or more channels for mobiles in its cell[‡]. The channels can be individual frequencies in FDMA, time slots in TDMA, or orthogonal codes or hopping patterns in case of spread-spectrum.

The paper first presents a framework for comparison of energy consumption due to MAC related activities. The activities considered are transmission and reception of a single packet and periodic packets. The average time the transmitter and the receiver are in use for each of the activities is determined through analysis and simulation. This framework is then applied to a set of protocols that includes IEEE 802.11 standard [8], Energy-conserving MAC (EC-MAC) [3], Packet reservation multiple access (PRMA) [9], Multiservices dynamic reservation - TDMA (MDR-TDMA) [4,5], and Distributed-queuing request update multiple access (DQRUMA) [6]. Of these, 802.11 was designed primarily for data traffic, PRMA was designed for voice and data traffic. The other three protocols are specifically designed to handle multimedia traffic. The results obtained from mathematical analysis are presented in the paper. These results have been validated through extensive discrete-event simulation.

The analysis here shows that protocols that aim to reduce the number of contentions perform better from an energy consumption perspective. IEEE 802.11, for example, senses the medium before transmitting. This results in less collisions than protocols using slotted ALOHA. The transmitter usage time is therefore also less than others and is almost independent of the traffic load. The receiver usage time of 802.11 tends to be higher than other protocols since the mobile may have to sense several slots before capturing the medium during heavy traffic. For PRMA, both the receiver and transmitter need to be powered up for the slotted ALOHA contention mode. As the traffic load increases, the mobile may encounter more and more collisions in trying to transmit a data packet. Thus both the receiver and transmitter usage times increase. Using short packets for contention in MDR-TDMA and DQRUMA reduces these usage times. However, in these systems there might be too many contentions during heavy traffic due to typical of slotted ALOHA protocols.

For messages with contiguous packets, our analysis shows that reservation ALOHA is more energy conservative than piggybacking. In DQRUMA, the explicit slot-by-slot announcement allows the base station to implement "optimal" and "just-in-time" scheduling. Because scheduling is done on a slot-by-slot basis, DQRUMA can potentially reduce packet latency. However, the burden placed on the receiver sub-system

[‡] Please note the term *cell* here is different from the term cell used to denote the basic transmission unit in ATM. The difference will be apparent based on the context.

to receive and decode during every slot weakens the protocol. Additionally, the transmitter needs to send a piggybacking request in each slot thus increasing the transmitter usage time. EC-MAC, which was specifically designed for energy efficiency, achieves better performance from power perspective compared to the other protocols. The energy consumption of EC-MAC is almost independent of the packet traffic load and the number of mobiles. During heavy system traffic load (both packet traffic and number of mobiles), the maximum energy consumption for EC-MAC is less than that for any other protocols under consideration here. One possible downside is the overhead due to the explicit transmission order transmitted during the request/update phase.

The rest of the paper is organized as follows. Section 2 describes some of the basic energy conservation principles partly addressed in [10]. Section 3 briefly describes the wireless access protocols studied in this paper. Section 4 provides the analysis of energy consumption of the studied protocols based on mathematical techniques. Section 5 presents the numerical results from both analytic models and simulation. Section 6 summarizes the paper.

2 Energy Conservation Principles

Mobile computers typically have limited energy for computing and communications because of the short battery lifetimes. Conserving battery energy in mobiles should be a crucial consideration in designing protocols for mobile computing. This issue should be considered through all layers of the protocol stack, including the application layer. Low-power design at the hardware layers uses different techniques including variable clock speed CPUs [11], flash memory [12], or disk spindowns [13]. At the application layer, low-power video compression [14], transcoding at the base station [15] and energy efficient database operation [16, 17] have been considered. In [18], the power drained by the network interface in hand-held devices was studied. An energy efficient probing scheme for error control in link layer is proposed in [19, 20]. The interaction of error control and forward error correction schemes in the link layer are studied from energy efficiency perspective in [21]. The problem of how a mobile in a wireless network should adjust its transmitter power so that the energy consumption is minimized has been considered in [22]. This paper is concerned with energy conservation in access layer protocol activities and recounts part of the discussion found in [10].

The chief sources of energy consumption in the mobile unit due to MAC related activities are the CPU, the transmitter, and the receiver. Mobile CPU usage may be reduced by relegating most of the high-complexity computation (related to media access) to the stationary network. Therefore, the focus of this work is on transceiver usage. The radio can operate in three modes: standby, receive, and transmit. In general, the radio consumes more power in the transmit mode than in the receive mode, and consumes least power in the standby mode. For example, the Proxim RangeLAN2 2.4 GHz 1.6 Mbps PCMCIA card requires 1.5W in transmit, 0.75W in receive, and 0.01W in standby mode [23]. In addition, turnaround between transmit and receive modes (and vice-versa) typically takes between 6 to 30 microseconds. Also, power consumption for Lucent's 15 dBm 2.4 GHz 2 Mbps Wavelan PCMCIA card is 1.82W in transmit mode, 1.80W in receive mode, and 0.18W in standby mode [24]. Similar figures are 3.0W, 1.48W, and 0.18W, respectively for a 24.5 dBm 915 MHz 2 Mbps PCMCIA card. The power consumption will be higher for higher bit rates due to the higher equalization complexity.

The objective of MAC protocol design should be to minimize energy consumption while maximizing protocol performance. The following are some principles that may be observed to conserve energy at the MAC level.

1. Collision should be eliminated as far as possible since it results in retransmissions leading to unnecessary energy consumption and also to possibly higher delay. Note that retransmissions cannot be completely avoided due to the high link error-rates and due to user mobility. For example, new users registering with the base station may have to use some form of random access protocol. However, using a small packet size for registration and bandwidth requests can reduce energy consumption.

   Techniques such as reservation and polling can help meet the requirement that collisions be minimized. Reservation and polling based protocols for wireless ATM networks have been proposed in [3,5] and [2,25], respectively.

2. In a typical wireless broadcast environment, the receiver has to be powered on at all times resulting in significant energy consumption. The receiver sub-system typically receives all packets and forwards only the packets destined for this mobile. For instance, this is the default mechanism used in IEEE 802.11 where the receiver is expected to keep track of channel status through constant monitoring.

   One possible way to reduce receiver power-on time is to broadcast a data transmission schedule for each mobile. This will enable a mobile to switch to standby mode until its alloted slots. This approach has been described in [3,26].

3. The IEEE 802.11 standard recommends the following technique for power conservation. A mobile that wishes to conserve power may switch to sleep mode and inform the base station of this decision. From that point on, the base station buffers packets destined for this mobile. The base station periodically transmits a beacon that contains information about such buffered packets. Upon waking up, the mobile listens for this beacon and informs the base station that it is ready to receive. The base station then forwards the buffered packets to the mobile station. This approach conserves power at the mobile but results in additional delays that may affect quality-of-service (QoS). It is essential to quantify this delay in the presence of QoS delay bounds for individual VCs.

4. The HIPERLAN standard for wireless LANs [27] provides two types of power saving mechanisms. The implicit mechanism turns on the equalizer only when the mobile is the intended destination of the downlink packet. The explicit mechanism allows the mobile to receive only during pre-arranged intervals instead of continuously. A mobile entering power-saver state informs a power-supporter mobile (possibly with a infrastructure power source) of the periodicity and length of the duration during which the mobile will be turned on. The p-supporter station receives and stores packets addressed for the power-saver mobiles it is supporting. This is essentially a distributed version of the 802.11 mechanism.

5. In the case of wireless ATM networks, if reservations are used to request bandwidth, it will be more efficient (power-wise and bandwidth-wise) to request multiple cells with a single reservation packet. For example, an ATM-aware MAC layer can request resources for a complete or partial AAL5 packet instead of a cell-by-cell basis. This suggests that the mobile should request larger chunks of bandwidth to reduce the reservation overhead leading to better bandwidth and energy consumption efficiency. Some of the earlier protocols utilize such reservation modes for CBR traffic. For more dynamic VBR traffic, occasional queue status updates are utilized to inform the base station of changing traffic needs [3,25].

6. Assume that mobiles transmit requests and that the base station uses a scheduling algorithm to allocate slots as in [3,5,6,25]. A distributed algorithm where each mobile computes the schedule independently may not be desirable because: (i) it may not receive all the reservation requests due to radio and error constraints, and (ii) schedule computation consumes energy and is thus better relegated to the base station. This suggests that a centralized scheduling mechanism will be more energy efficient.

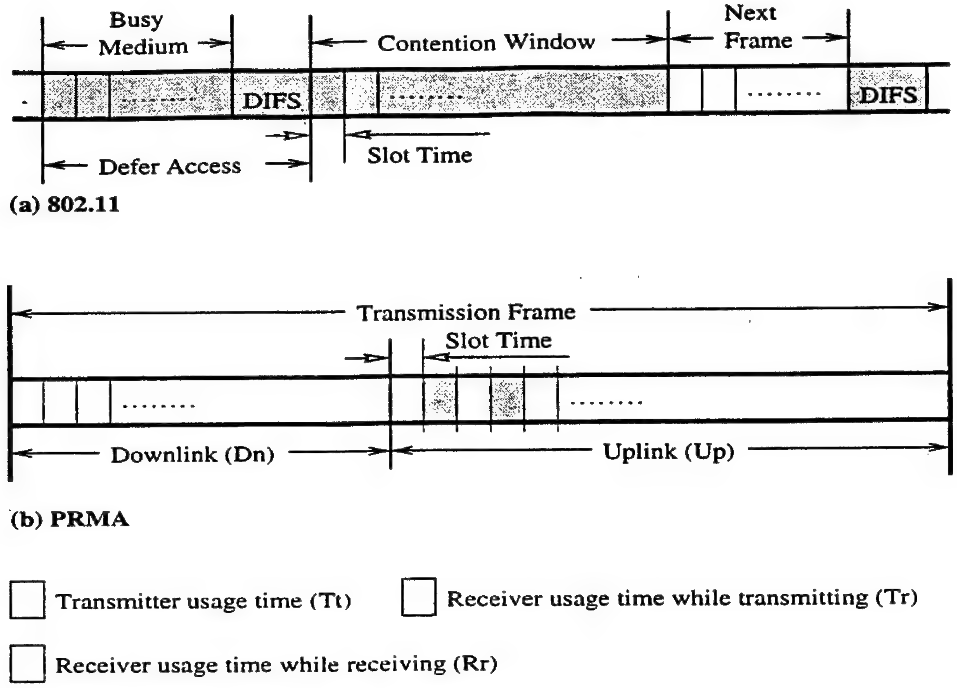
The principles discussed above have been derived from different access protocols and are by no means an exhaustive list of efficient energy utilization techniques for the MAC protocol.

3 MAC Protocols

This section briefly describes the wireless access protocols studied in this paper. Figs. 1–2 show the channel access methods for these protocols.

The IEEE 802.11 standard [8] for wireless LANs defines multiple access using a technique based on Carrier Sense Multiple Access / Collision Avoidance (CSMA/CA). The basic access method is the Distributed Coordination Function (DCF) shown in fig. 1(a). A backlogged mobile may immediately transmit packets when it detects free medium for greater than or equal to a DIFS (DCF Interframe Space) period. If the carrier is busy, the mobile defers transmission and enters the backoff state. The time period following the unsuccessful transmission is called the contention window and consists of a pre-determined number of slots. The mobile, which has entered backoff, randomly selects a slot in the contention window, and continuously senses the medium during the time up to its selected contention slot. If it detects transmission from some other mobiles during this time period, it enters the backoff state again. If no transmission is detected, the mobile transmits the access packet and captures the medium. Extensions to the basic protocol include providing MAC-level acknowledgments and ready-to-send (RTS) and clear-to-send (CTS) mechanisms.

Packet reservation multiple access (PRMA) [9] was proposed for integrating voice and data traffic. The PRMA system is closely related to reservation ALOHA since it merges characteristics of slotted ALOHA and TDMA protocols. Packets in PRMA are grouped into periodic information and random information packets. Once a mobile with periodic information transmits successfully a packet in an available slot, that slot in future frames can be reserved for this mobile. However, mobiles with random information need to



(a) 802.11

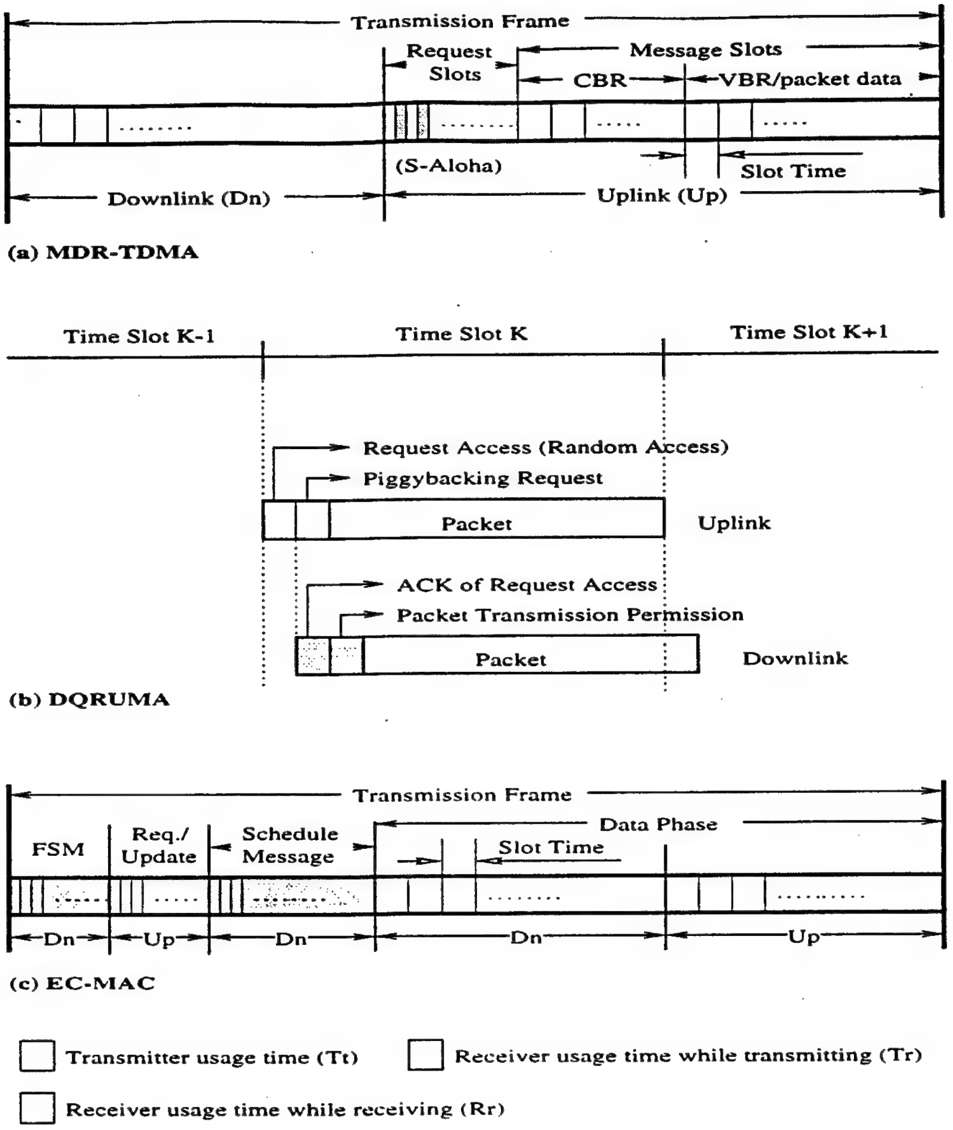
(b) PRMA

Figure 1: Channel access methods for IEEE 802.11 and PRMA.

contend for an available slot each time. The protocol is depicted in fig. 1(b).

The multiservices dynamic reservation TDMA protocol (MDR-TDMA) [4], shown in fig. 2(a) supports CBR, VBR, and ABR traffic by dividing TDMA frames for different types of traffic and allocating them dynamically. The TDMA frame is subdivided into N_r request slots and N_t message slots. Each message slot provides for transmission of a packet or an ATM-like *cell*. Request slots are comparatively short and are used for initial access in slotted ALOHA contention mode. Of the N_t message slots, a maximum of $N_v < N_t$ slots in each frame can be assigned for CBR voice traffic. VBR and packet data messages are dynamically assigned one or more 48-byte slots in the TDMA interval following the last allocated voice slot in a frame. The basic channel access scheme follows a combination of circuit mode reservation of slots over multiple TDMA frames for CBR voice calls with dynamic assignment of remaining capacity for VBR or packet data traffic. In addition to first-come-first-served (FCFS) scheduling, time-of-expiry (TOE) approach has been studied to improve delay performance of real-time data traffic. Energy efficiency issues, however, are not specifically addressed in the protocol definition.

The distributed-queuing request update multiple access (DQRUMA) protocol [6] is shown in fig. 2(b). The base station employs a random access protocol and packet scheduling policy based on traffic and service requirements. Mobiles send a transmission request only when packet(s) join an empty queue. All subsequent packets that arrive at the queue can piggyback transmission requests. Two request access protocols have been studied: the ALOHA random access protocol, and a generalization of the Binary Stack Algorithm. The scheduling policy considered is a round-robin packet transmission policy. Since the slots are scheduled on a finer grain in DQRUMA, the requirement that the mobile should listen during every slot places a high



(a) MDR-TDMA

(b) DQRUMA

(c) EC-MAC

Figure 2: Channel access methods for MDR-TDMA, DQRUMA, and EC-MAC.

burden on the mobile's power resources.

The protocol design of energy-conserving medium access control (EC-MAC) protocol [3] is driven by energy consumption, diverse traffic type support, and QoS support considerations. The protocol is defined using fixed-length frames since each mobile receiver will precisely know the time of the next beacon transmission. This enables the receiver to power off knowing precisely when the next frame will start. The frame is divided into multiple phases: frame synchronization, request/update phase, new-user phase, schedule message, and data phase. At the start of each frame, the base station transmits the frame synchronization message (FSM) on the downlink. This message contains framing and synchronization information, the uplink transmission order for reservations, and the number of slots in the new user phase. The request/update phase is composed of uplink request transmissions from the mobiles. During this phase, each registered mobile transmits new connection requests and queue status of established queues according to the transmission order. The base station then broadcasts the transmission schedule for the data phase using a schedule message. Mobiles receive the broadcast and power on the transmitters and receivers at the appropriate time. The new-user phase allows new mobiles that have entered the cell coverage area to register with the base station. The comparison analysis in next section assumes that all mobiles in the cell coverage area have already registered with the base station. Fig. 2(c), therefore, does not incorporate the new-user phase.

A number of other access protocols for wireless multimedia networks based on ATM have been proposed in the literature, some of which are summarized in [7]. The protocols described here are chosen to represent the major categories of multiple access protocols for local area wireless networks.

4 Energy Consumption Analysis

This section analyzes the energy consumption during two important protocol activities at the mobile's MAC layer: packet transmission and reception. The analysis assumes that the mobile has already registered with the base station. All the mobile transmissions are directed to and all mobile receptions are received from the base station. For each of the two activities, we quantify the time spent utilizing the transmitter and receiver. For networks with power control, the energy spent for transceiver is varying in time which depends on the power management schemes. Since the MAC protocols studied here are not specified for any particular power control schemes, we assume each protocol uses the fixed transmitted power. Hence, energy consumed is proportional to the amount of time spent utilizing each resource. For simplicity, we assume collisions result from packet errors only rather than noise or interference. A discussion of the interference and noise problems can be found in [28]. For transmitting packets (either single or periodic packets), T_r and T_t are defined as the average time spent using the receiver and transmitter, respectively. For receiving packet(s), the average receiver usage time is given by R_r.

Table 1 summarizes the system parameters and definitions used in the analysis. We assume that time is slotted and the time necessary to receive or transmit a packet is L units of time, where L denotes the length of a data packet. A reservation or contention packet is used to gain access to the medium, and its length is taken to be l units of time. The parameter a is the time spent decoding a slot while the mobile listens to the downlink for the packet destined to it. The average number of slots each mobile needs to sense before receiving a packet destined to it is defined as X. The system contains N mobiles. The analysis is based on how much energy a mobile needs for transmitting/receiving a packet or packets while there are other C contending mobile terminals with packet arrival rate λ. Λ is the total transmission rate of newly generated plus retransmitted packets ($\Lambda \geq \lambda$). G is defined as the offered traffic load ($G = \Lambda L$). $E[L_t]$ is the average time to receive/transmit voice talkspurts. For IEEE 802.11, K is the size of contention window.

Name	Description
T_t	Average transmitter usage time while transmitting packet(s)
$.T_r$	Average receiver usage time while transmitting packet(s)
R_r	Average receiver usage time while receiving packet(s)
L	Time to receive/transmit a packet
l	Time to receive/transmit a reservation/contention packet
a	Time spent decoding a slot
X	Number of slots sensed before receiving the packet destined to it
N	Number of mobile stations ($N > 0$)
C	Number of other contention stations ($C = N - 1$)
λ	Packet arrival rate per each mobile
Λ	Total transmission rate per mobile (newly generated + retransmitted)
G	Offered traffic load ($G = \Lambda L$)
$E[L_t]$	Time to receive/transmit voice talkspurts
T_s	Transmitter/receiver usage time for a successful contention
T_f	Transmitter/receiver usage time for a failure contention
p	Probability of a failure contention
P_s	Probability of a successful contention (802.11)
P_f	Probability of a failure contention (802.11)
K	Size of contention window (802.11)
L_A	Length of an acknowledgment packet (PRMA)
δ	Number of transmission permission issued (DQRUMA)
Δ	Queue length (DQRUMA)

Table 1: System Parameters

4.1 802.11

During packet transmission in 802.11, the mobile needs to listen to the medium until it is free. Fig. 1(a) indicates that the receiver is the most utilized resource. If the medium is active, the average time spent using the receiver is:

$$T_r = E\left[L_1\right] + E\left[\tau_1\right] \tag{1}$$

where $E[L_1]$ is the *expected value* (or *average* or *mean value*) of time the receiver is turned on when some other mobile is currently transmitting its data packet. $E[\tau_1]$ is the expected value of time spent using the receiver when this mobile stays in backoff procedure due to unsuccessful contention before capturing the medium. $E[L_1]$ can be obtained by:

$$E\left[L_1\right] = \frac{L}{2} + DIFS \tag{2}$$

To evaluate $E[\tau_1]$, define the probability that some other mobiles transmit in the contention window before

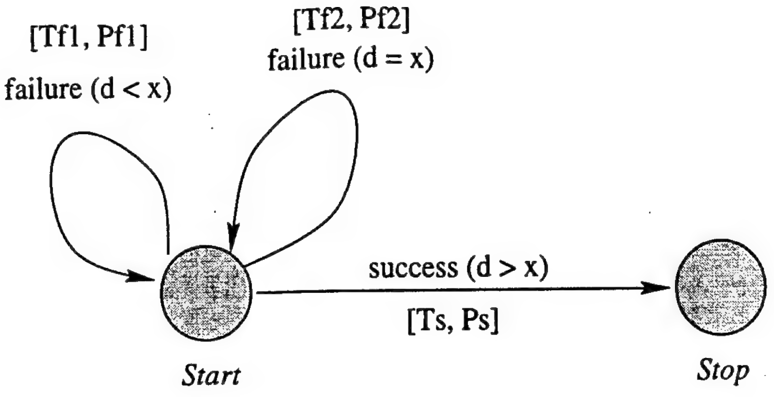


Figure 3: Regenerative model used for T_r in 802.11.

this mobile does is P_{f1}, and the corresponding average time the receiver is utilized is T_{f1}. The probability that two or more mobiles sense the medium idle for a sufficient period of time, and attempt transmissions simultaneously is P_{f2}, and the corresponding average time the receiver is utilized is T_{f2}. The probability that it contends successfully is P_s, and the corresponding time receiver is turned on is T_s. Using *regenerative method* [29] to obtain $E[\tau_1]$ as follows (fig. 3):

$$E\left[\tau_1\right] = P_{f1}\left(T_{f1} + E\left[\tau_1\right]\right) + P_{f2}\left(T_{f2} + E\left[\tau_1\right]\right) + P_s T_s \tag{3}$$

Solving equation (3) for $E[\tau_1]$ gives

$$E\left[\tau_1\right] = \frac{P_{f1}T_{f1} + P_{f2}T_{f2} + P_s T_s}{\left(1 - P_{f1} - P_{f2}\right)} \tag{4}$$

Let x be the slot that this mobile randomly chooses in the contention window, where $1 \leq x \leq K$ (Recall that K is the size of the contention window). Let d be the smallest slot other mobiles choose. If no one transmits in slots before x, i.e. $d > x$, this mobile captures the medium and transmits its packet. If, on the other hand, the mobile detects transmission from other mobiles in time slot d, where $d < x$, it enters the backoff state again. When one or more mobiles attempt transmission simultaneously as this mobile does, d then equals x. To obtain P_{f1}, P_{f2}, and P_s, we calculate the probability (P_f) that $d \leq x$ first.

$$P_f = \sum_{x=1}^{K} \frac{1}{K} \sum_{m=1}^{C} \left[\binom{C}{m}\left(1 - e^{-G}\right)^m \left(e^{-G}\right)^{C-m}\right]\left[1 - \left(\frac{K-x}{K}\right)^m\right] \tag{5}$$

In equation (5), the first term $[\binom{C}{m}(1 - e^{-G})^m(e^{-G})^{C-m}]$ represents the probability that some other mobile (or mobiles) also generates packet(s) before contention window begins. Some of the packets that arrive in the duration L before the contention window will have to enter the backoff procedure due to unsuccessful contention. As mentioned earlier, there are other C contending mobile stations, with the arrival of packets at each mobile as a Poisson process with rate λ. Let Λ ($\Lambda \geq \lambda$) be the rate of packet attempting transmission over the channel per user. This includes newly generated plus retransmitted packets. Following the analysis in [30, 31], we assume that the composite message generation per user is Poisson distributed. Let G be the average number of total arrivals in the duration of L. Therefore, $G = \Lambda L$. The probability that a mobile is active during time interval L is then $(1 - e^{-G})$. The probability that m over C mobiles are active can be obtained by binomial distribution as above. The second term in equation (5) represents the probability some other mobile (or mobiles) chooses a slot d where $d \leq x$ thereby causing the mobile to enter backoff state again. $\left(\frac{K-x}{K}\right)^m$ is the probability that all other mobiles choose the slot after x. Please note equation (5) holds for $C > 0$. When $C = 0$, P_f equals 0.

The probability that some other mobiles transmit in the contention window before this mobile does, i.e. $d < x$, equals 0 when $C = 0$. When $C > 0$, it is

$$P_{f1} = \sum_{x=1}^{K} \frac{1}{K} \sum_{m=1}^{C} \left[\binom{C}{m}\left(1-e^{-G}\right)^{m}\left(e^{-G}\right)^{C-m}\right]\left[1-\left(\frac{K-x+1}{K}\right)^{m}\right] \tag{6}$$

From equations (5) and (6), P_{f2} equals $(P_f - P_{f1})$. P_s can be obtained as the probability that there are no other arrivals at the other mobiles plus the probability that every other mobile where packets arrive chooses slot greater than x. Thus,

$$P_s = \left[\binom{C}{0}\left(1-e^{-G}\right)^{0}\left(e^{-G}\right)^{C}\right] + \sum_{x=1}^{K} \frac{1}{K} \sum_{m=1}^{C} \left[\binom{C}{m}\left(1-e^{-G}\right)^{m}\left(e^{-G}\right)^{C-m}\right]\left[\left(\frac{K-x}{K}\right)^{m}\right] \tag{7}$$

P_s can be calculated as $(1 - P_{f1} - P_{f2})$ or $(1 - P_f)$ as well.

If no one transmits in slots before x, this mobile captures the medium and transmits its packet. There are C other mobiles. Let A be the number of active mobiles among these C mobiles. T_s can be obtained by

$$T_s = \sum_{x=1}^{K} x \left[\frac{(K-x)^{A}}{\sum_{x'=1}^{K}(K-x')^{A}}\right] \tag{8}$$

where

$$A = \sum_{m=1}^{C} m \left[\binom{C}{m}\left(1-e^{-G}\right)^{m}\left(e^{-G}\right)^{C-m}\right] \tag{9}$$

For a given x, there are $(K-x)^{A}$ possibilities that other A active mobiles choose their slots greater than x. The sample space is the sum of them. The second term in equation (8), therefore, is the probability of each given x. Equation (8) shows that the length of T_s depends on both C and G, the number of mobiles and the traffic load.

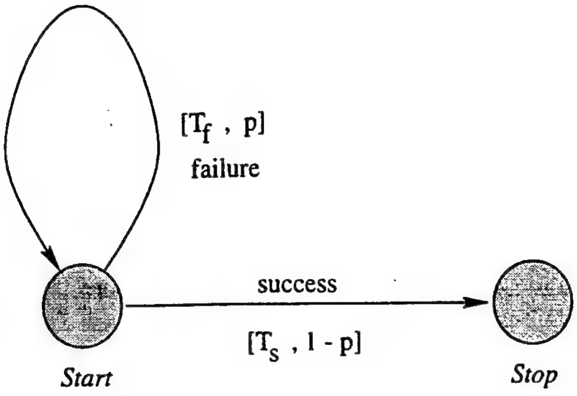


Figure 4: Regenerative model used for 802.11 (except for computing T_r), PRMA, MDR-TDMA, and DQRUMA.

As defined previously, d is the smallest slot other mobiles choose. If $d < x$, this mobile enters the backoff state. We can estimated T_{f1} by

$$T_{f1} = L + DIFS + \sum_{d=1}^{K} d \left[\frac{(K-d)\left[\sum_{i=1}^{A}\binom{A}{i}(1)^{i}(K-d)^{A-i}\right]}{\sum_{d'=1}^{K}(K-d')\left[\sum_{i=1}^{A}\binom{A}{i}(1)^{i}(K-d')^{A-i}\right]} \right] \tag{10}$$

For a given d, there are $(K-d)$ possibilities for x. For other A active mobiles, it might be possible that more than one mobile chooses the same d which is less than x. Therefore, there might be 1 to A mobile(s) choose(s) the same time slot d, and the rest of them choose the time slot greater than d. For a given d, hence, there are $\left[\sum_{i=1}^{A}\binom{A}{i}(1)^{i}(K-d)^{A-i}\right]$ possibilities for other A active mobiles. Therefore, d can be estimated by the third term in equation (10). Please note $(1)^{i}$ here can be eliminated. We reserve it for easier understanding.

When one or more other active mobiles attempt transmission simultaneously as this mobile does, d then equals x. For each given x, there is at least one mobile which has $d = x$. Consequently,

$$T_{f2} = L + DIFS + \sum_{x=1}^{K} x \left[\frac{(K-x+1)^{A} - (K-x)^{A}}{\sum_{x'=1}^{K}(K-x'+1)^{A} - (K-x')^{A}} \right] \tag{11}$$

By replacing T_{f1}, T_{f2}, T_s, P_{f1}, P_{f2}, and P_s in equation (4), we can get $E[\tau_1]$. T_r in equation (1) can then be evaluated by equations (2) and (4).

During the backoff period, the transmitter is not used most of the time. The transmitter is utilized only when the mobile captures the channel or when one or more other mobiles sense the medium idle for a sufficient period of time and attempt transmissions simultaneously, i.e. $d = x$. This will result in collision and will be resolved using backoff techniques. Assume the mobile detects the collision after one slot time. The transmitter usage time is given by

$$T_t = E[\tau_2] \tag{12}$$

Regenerative method is used to obtain $E[\tau_2]$ as follows (fig. 4):

$$E[\tau_2] = p(T_f + E[\tau_2]) + (1-p)\,T_s \tag{13}$$

Solving equation (13) for $E[\tau_2]$ gives

$$E[\tau_2] = \frac{p(T_f - T_s) + T_s}{1-p} \tag{14}$$

where $T_f = 1, T_s = L$, and $p = P_{f2}$. T_t in equation (12), therefore, is obtained by equation (14).

During packet reception, the receiver has to be turned on during the entire downlink transmission. It reads the header of every downlink packet, and moves to standby mode if the packet is not destined for it. If the receiver senses X slots and a is the time spent decoding each slot, the receiver usage time is given by $R_r = aX + L$. Let S be the probability that the receiver senses this slot is destined to it. It is reasonable to assume that destinations of packets sent by the base station are uniformly distributed over all the mobiles in the cell. For N mobiles in the cell, S equals $\frac{1}{N}$. The expected number of slots a mobile has to receive before its intended packet is then obtained by

$$E[X] = N \tag{15}$$

Therefore,

$$R_r = aN + L \tag{16}$$

The analysis above is based on the transmitting and receiving of data packets. Since the 802.11 standard does not detail the handling of voice traffic, we ignore voice packets in our analysis of 802.11.

4.2 PRMA

The PRMA [9] system is closely related to reservation ALOHA, since it merges characteristics of slotted ALOHA and TDMA protocols. During packet transmission, both the transmitter and receiver are utilized. The mobile transmits its packet in the next slot after the packet is generated. If two or more mobiles transmit simultaneously in the same slot, collision results. It continues to transmit its packet until the base station acknowledges successful reception of the packet. As discussed above, L denotes the length of a data packet. Let L_A be the length of an acknowledgment. By applying the regenerative model, the average time spent

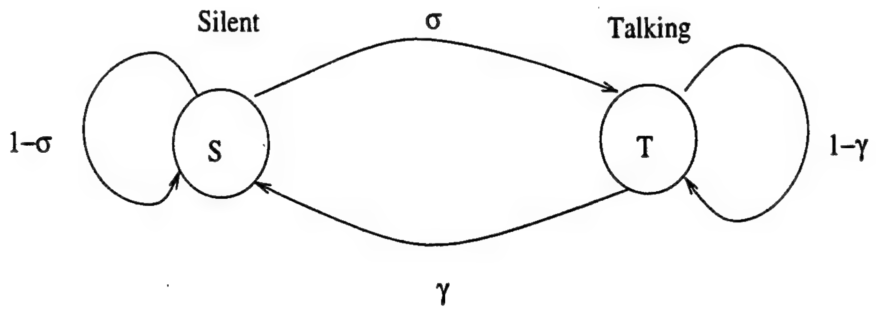


Figure 5: Two-state Markov process based voice model, representing a source with speech activity detector.

using the transmitter can be obtained by replacing $T_f = T_s = L$ in equation (14):

$$T_t = \frac{L}{1-p} \tag{17}$$

In slotted ALOHA, all packets arriving during previous slot are transmitted together in current slot. Therefore, p is evaluated as follows:

$$p = \sum_{m=1}^{C} \binom{C}{m} \left(1 - e^{-G}\right)^{m} \left(e^{-G}\right)^{C-m}, \quad C > 0 \tag{18}$$

where G is as defined previously. As discussed in 802.11, we assume that the composite of newly generated plus retransmitted packets is Poisson distributed [30, 31]. In equation (18), $p = 0$ if $C = 0$. Similarly, the average time spent using the receiver is:

$$T_r = \frac{L_A}{1-p} \tag{19}$$

During packet reception, the receiver has to be turned on during the entire downlink transmission to decode the intended receiver information. As discussed for 802.11, the receiver usage time is:

$$R_r = aN + L \tag{20}$$

The analysis above is based on the transmitting and receiving of one single packet. Suppose there are two different kinds of packets: data packet and voice packet. If each data packet needs to contend for transmission, T_t, T_r, and R_r for a data packet are same as those in equations (17), (19), and (20), respectively.

Voice packet, however, may reserve the same time slot in future frames until the end of talkspurts. Only the first packet needs to contend by sensing the medium. Voice traffic is modeled as a two-state Markov process (fig. 5) representing a source with a *slow speech activity detector* (SAD) [32]. The probability that a principal talkspurt with mean duration t_1 seconds ends in a frame of duration t is

$$\gamma = 1 - e^{-t/t_1} \tag{21}$$

The probability that a silent gap with mean duration t_2 seconds ends in a frame of duration t is

$$\sigma = 1 - e^{-t/t_2} \tag{22}$$

Thus γ is the probability that a source makes a transition from talkspurt state to silent state and σ is the probability that the source makes a transition from silent state to talkspurt state. If a voice source generates one voice packet in each frame, a talkspurt of t_1 seconds contains $\frac{t_1}{t}$ packets. Therefore, a talkspurt needs $\frac{t_1 L}{t}$ units of time to be transmitted. At the end of a talkspurt, another talkspurt may follow with probability $1 - \gamma$, or the source may go silent with probability γ. Let $E[L_t]$ denote the expected value of time spent using the transmitter until the silent gap begins. $E[L_t]$ can be obtained by equation (14) by applying the regenerative model, where $p = 1 - \gamma$, $T_f = \frac{t_1 L}{t}$, and $T_s = 0$. Therefore,

$$E\left[L_t\right] = \frac{t_1 L}{t}\left(\frac{e^{-t/t_1}}{1 - e^{-t/t_1}}\right) \tag{23}$$

We then get T_t and R_r for *talkspurts* as follows:

$$T_t = \frac{L}{1-p} + E\left[L_t\right] - L \tag{24}$$

$$R_r = aN + E\left[L_t\right] \tag{25}$$

where p and $E[L_t]$ can be obtained by equation (18) and (23), respectively. For voice packets, T_t is, in other words, equal to the average time it takes to transmit the first packet using contention $\left(\frac{L}{1-p}\right)$ plus the average time to transmit the rest of the talkspurts ($E[L_t] - L$). Once the first packet has successfully gained access to the medium, the receiver does not need to listen to the channel for the rest of the talkspurt(s). The subsequent packets in the talkspurt(s) will be allocated the same slot in the following frames. Thus, T_r for *talkspurts* is same as that in equations (19).

4.3 MDR-TDMA

MDR-TDMA [4] divides TDMA frames for different types of traffic and allocates them dynamically. The TDMA frame is subdivided into N_r request slots and N_t message slots. Each message slot is used for the transmission of a packet. Request slots on the other hand are comparatively short and are used for initial access using slotted ALOHA contention mode. Of the N_t message slots, a maximum of $N_v < N_t$ slots in each frame can be assigned for CBR voice traffic. Other packets are dynamically assigned in the TDMA

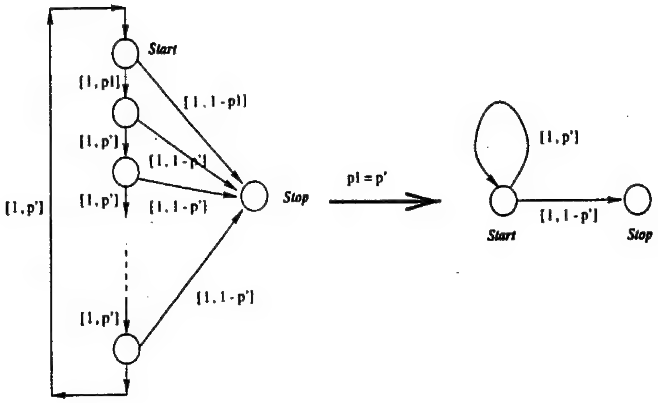


Figure 6: Models for MDR-TDMA.

interval following the last allocated voice slot in a frame. The frame structure is defined in fig. 2(a).

Let l denote the length of a contention packet in request slots and the length of an acknowledgment. In slotted ALOHA, all packets arriving in previous slot will be transmitted together in current slot. If packets are generated in the duration N_r and N_t, the probability that the first contention packet in N_r contends unsuccessfully is denoted by p_1. The probability p_1 is computed using equation (18) for $G_1 = \Lambda L N_t$. Other contention packets in N_r have the probability p', for $G = \Lambda l$. If all mobiles generate and retransmit packets only in N_r, $p_1 = p'$ (fig. 6). By normalizing the contention period from all slots in the frame to slots in N_r only, we can use the regenerative model. The average time spent using the transmitter can be obtained by equation (14):

$$T_t = \frac{l}{1-p} + L \tag{26}$$

where p can be obtained by equation (18). The average time spent using the receiver is:

$$T_r = \frac{l}{1-p} \tag{27}$$

During packet reception, the receiver has to be turned on during the entire downlink transmission to decode the intended receiver information. As discussed for 802.11, the receiver usage time is

$$R_r = aN + L \tag{28}$$

The analysis above is valid for a single packet and for a data packet if data packets need to contend for an available slot each time. However, once a mobile transmits successfully a voice packet in an available slot, that slot in future frames can be reserved for this mobile until the end of talkspurts. By using the same model in PRMA, we then get T_t and R_r for *talkspurts* as follows:

$$T_t = \frac{l}{1-p} + E\left[L_t\right] \tag{29}$$

$$R_r = aN + E\left[L_t\right] \tag{30}$$

where $E[L_t]$ can be obtained by equation (23). T_r for *talkspurts* is same as that in equation (27).

4.4 DQRUMA

In DQRUMA [6], mobile users send transmission requests during a request-access (RA) subslot of every slot or requests are piggybacked on to current data transmissions. Scheduling is done on a slot-by-slot basis and an explicit announcement at the beginning of each slot identifies the "owner" of next slot. Access during the RA subslot is accomplished using a random access mechanism.

To transmit a packet, the initial request is sent using slotted ALOHA. The acknowledgment of successful reservation receipt may follow in the subsequent slot (depending on propagation delay). The mobile receiver has to be powered on for reception of this acknowledgment. Subsequent reservations may be piggybacked on to outgoing data packets. After the reservation is received, the receiver has to receive the downlink allocation information for every subsequent slot until the mobile is allocated transmission permission. The timing diagram of DQRUMA is depicted in fig. 2(b).

Let L denote the length of a data packet as before. Let l be the length of packets for RA, piggybacking, and transmission permission. By applying the regenerative model, the average time spent using the transmitter can be obtained by equation (14):

$$T_t = \frac{l}{1-p} + L \tag{31}$$

where p can be obtained by equation (18). Similarly, the average time spent using the receiver is:

$$T_r = \frac{l}{1-p} + \delta l \tag{32}$$

where δl is the average time while the receiver is utilized for transmission permissions. The value of δ depends on the scheduling algorithm executed in the base station.

To achieve downlink packet reception, the receiver has to be turned on during the beginning of each slot to decode the intended receiver information. As the discussion for 802.11, the receiver usage during reception is

$$R_r = aN + L \tag{33}$$

The analysis above is for the initial request packet. Once a mobile transmits the initial packet successfully, subsequent packets are requested by piggybacking until the queue is empty. Both data and voice packets are transmitted by this method in DQRUMA. T_t, T_r, and R_r can be obtained by following equations:

$$T_t = \frac{l}{1-p} + L + (\Delta - 1)\,(l + L) \tag{34}$$

$$T_r = \frac{l}{1-p} + \Delta\delta l \tag{35}$$

$$R_r = \Delta\ (aN + L) \tag{36}$$

where the value of Δ depends on the queue length. For voice talkspurts, Δ equals $E[L_t]$ in equation (23). However. δ depends on the scheduling algorithm executed in the base station.

4.5 EC-MAC

In EC-MAC [3], once a mobile gets admission to this cell coverage area using new-user phase, it listens to the downlink of FSM for the transmission order. The mobile then sends out new connection requests and queue status of established queues by uplink in request/update phase to the base station. The base station schedules the requests from mobiles based on the traffic types and QoS, and then broadcasts the schedule that contains the slot allocations for the subsequent data phase. The data phase includes downlink transmissions from the base station, and uplink transmissions from the mobiles. Mobiles, therefore, send out transmission requests and data traffic without collision after they have registered with the base station. This significantly reduces the energy consumption.

Let l be the length of packets used during FSM, request/update, and schedule phases. Mobile first listens to transmission order, and then sends out its request/update. After that, the mobile sends its packet in the data phase during its scheduled time. Therefore,

$$T_t = l + L \tag{37}$$

In the FSM phase, mobile listens to downlink until it gets transmission order. The maximum time spent using the receiver is $l\ \eta$, where η is the maximum number of downlink transmission. Similarly, the maximum time the receiver is utilized during schedule reception is $l\ \psi$, where ψ is the maximum number of permissions in the schedule phase. Note that the assumption here is that the downlink in request/update and schedule phases are long enough to accommodate all mobiles. Simulation studies in [3] show that the assumption is rational. The expected time receiver is turned on for sending a packet is given by:

$$2\,l \leq T_r \leq l\ (\eta + \psi) \tag{38}$$

To achieve downlink packet reception, the receiver has to be turned on during the schedule message. After the mobile gets the schedule, it powers on its receiver at the appropriate time in data phase. Let ψ be the maximum number of schedule beacons as discussed above.

$$l + L \leq R_r \leq l\,\psi + L \tag{39}$$

Note that the mobile only needs to listen to the schedule message once to determine its allocated slots in both uplink and downlink parts of the data phase. Therefore, equations (38) and (39) could be reduced to one of two possibilities: either R_r remains the same and T_r is reduced to

$$l \leq T_r \leq \eta\, l \tag{40}$$

or else, T_r is the same as (38) but R_r is equal to L.

The analysis above is valid for a single packet and for a data packet if data packets need to contend for an available slot each time. However, once a mobile successfully transmits a voice packet in an available slot, that slot in future frames may be reserved for this mobile until the end of the talkspurts. Using the same voice model as in PRMA, we get T_t and R_r for *talkspurts* as follows:

$$T_t = l + E\,[L_t] \tag{41}$$

$$l + E\,[L_t] \leq R_r \leq l\,\psi + E\,[L_t] \tag{42}$$

where $E[L_t]$ can be obtained by equation (23). T_r for *talkspurts* is same as that in equation (38).

5 Numerical Results

This section provides the numerical results for the comparison presented in last section. Figures for a single packet and periodic packets (voice talkspurts) are presented. Energy consumption for a Proxim RangeLAN2 radio card is also provided. The results are obtained for a channel transmission rate of 2 Mbps. Voice traffic is coded with 32 Kbps. The length of a packet (L) is 64 bytes. We assume that only 56 bytes are *useful* data

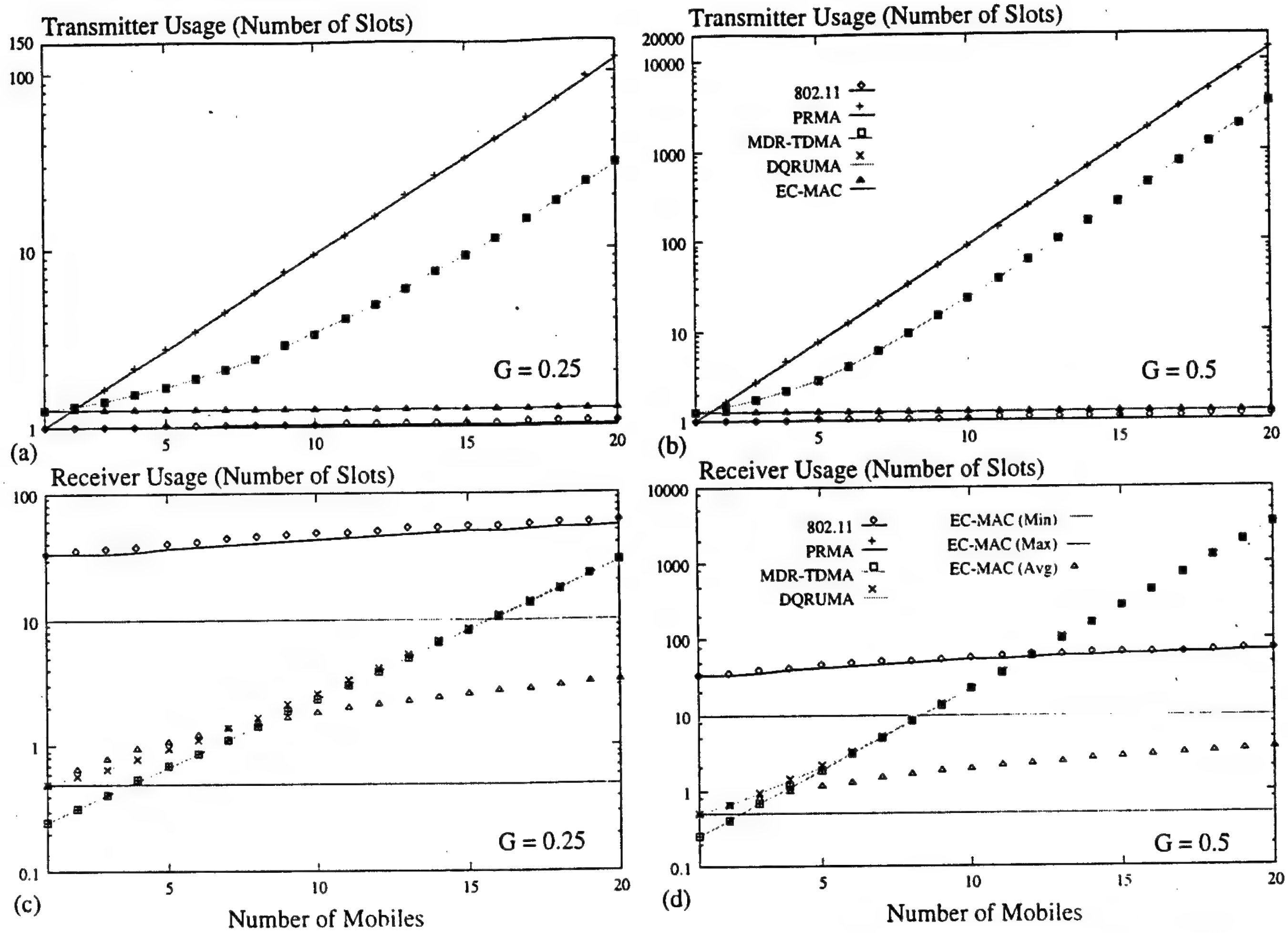


Figure 7: (a)-(b) Transmitter usage time (T_t), and (c)-(d) receiver usage time (T_r) versus number of mobiles (N) for transmitting a single packet. G is offered traffic load including newly generated and retransmitted packets. The figures are plotted for $G \in \{0.25, 0.5\}$. Points represent simulation and lines represent analysis.

after all coding schemes, header fields, error checksums, etc. are considered. The length of a contention packet (l) and acknowledgment (L_A) is 16 bytes. One slot time is 0.256 ms and length of slot is 64 bytes as well. For 802.11, the size of the contention window (K) is 64. The values of DIFS in 802.11 standard are 0.128 ms and 0.052 ms for frequency hopping spread spectrum (FHSS) and direct sequence spread spectrum (DSSS), respectively. Although figures with DIFS in FHSS are not shown, the results are almost identical to those in DSSS. Note that the Proxim radio has been used merely to obtain typical energy consumption values. It does not imply that all these access protocols can be implemented on a Proxim card. The results should therefore be construed as merely indicative of the performance trends.

The proposed analytic models are validated by extensive discrete-event simulation. Simulation results have been obtained using the stochastic self-driven discrete-event models, written in C with YACSIM [33]. Simulation was done with the same assumptions set in section 4. YACSIM is a C based library of routines

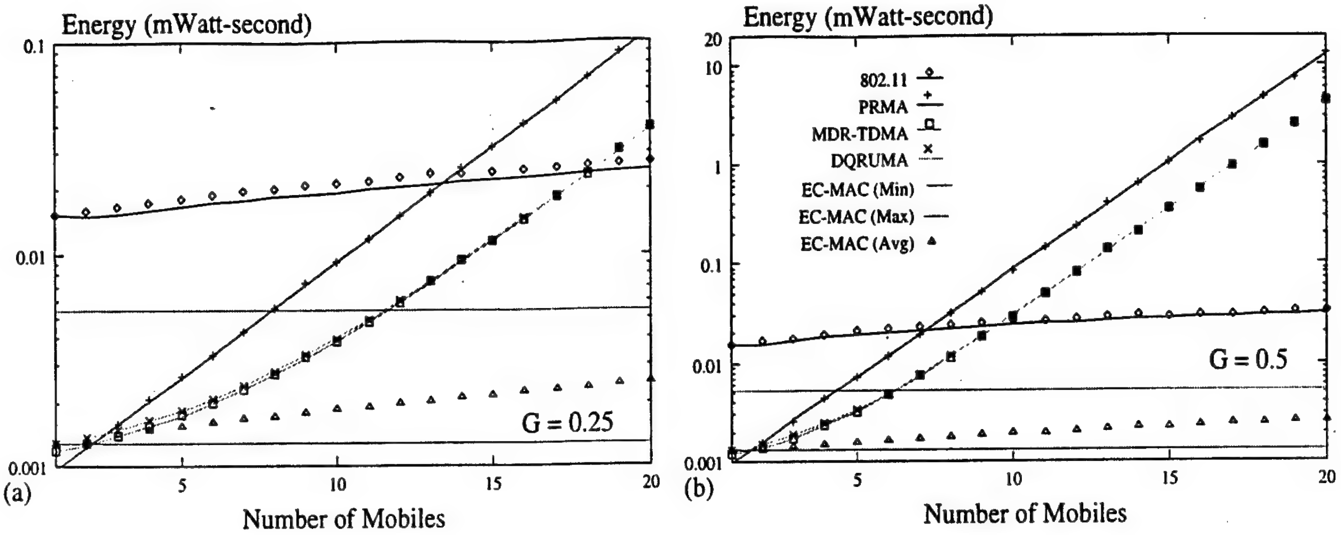


Figure 8: Energy spent per useful bit transmitted using Proxim's RangeLAN2 2.4 GHz radio, versus number of mobiles for transmitting a single packet. G is offered traffic load including newly generated and retransmitted packets. The figures are plotted for $G \in \{0.25, 0.5\}$. Points represent simulation and lines represent analysis.

that provides discrete-event and random variate facilities. Steady state transaction times were measured. Simulation convergence was obtained through the replication/deletion method [34] with a 97% confidence in a less than 5% variation from the mean. In figs. 7– 10, results from analysis and simulation are presented by lines and points, respectively.

Fig. 7 shows the transmitter and receiver usage times while transmitting a single packet. For 802.11, the mobile senses the medium before attempting to transmit. Collision occurs only when two or more mobiles choose the same slot in the contention window. The mobile transmits its packet after it captures the medium successfully in the contention window. Hence, fig. 7 (a) and (b) indicate that the transmitter usage time is almost independent of the number of mobiles. However, the probability that the mobile under consideration contends successfully decreases slightly as the traffic load increases. Fig. 7 (c) and (d) indicate that the receiver usage time increases as the number of mobiles increases. Since the receiver is the most utilized resource in 802.11, fig. 7 shows the T_r in 802.11 is larger than others when the traffic load is light. T_t, on the other hand, is much less than other protocols.

For PRMA, both receiver and transmitter need to be powered on in the slotted ALOHA contention mode. The transmitter is utilized for a packet transmission duration and the receiver is turned on to receive the acknowledgment. As the traffic load increases, the packet may suffer more collisions. Therefore, both the receiver and transmitter usage times increase. MDR-TDMA and DQRUMA also use slotted ALOHA to contend for a channel, but they employ a much shorter packet length. Hence, the two protocols have the same characteristics as PRMA does except that the time usage is less. In fig. 7 (a) and (b), MDR-TDMA and DQRUMA have the same transmitter usage time. Because reservation ALOHA is used in MDR-TDMA, packets in MDR-TDMA knows which slot to transmit after the initial contention. In DQRUMA, however,

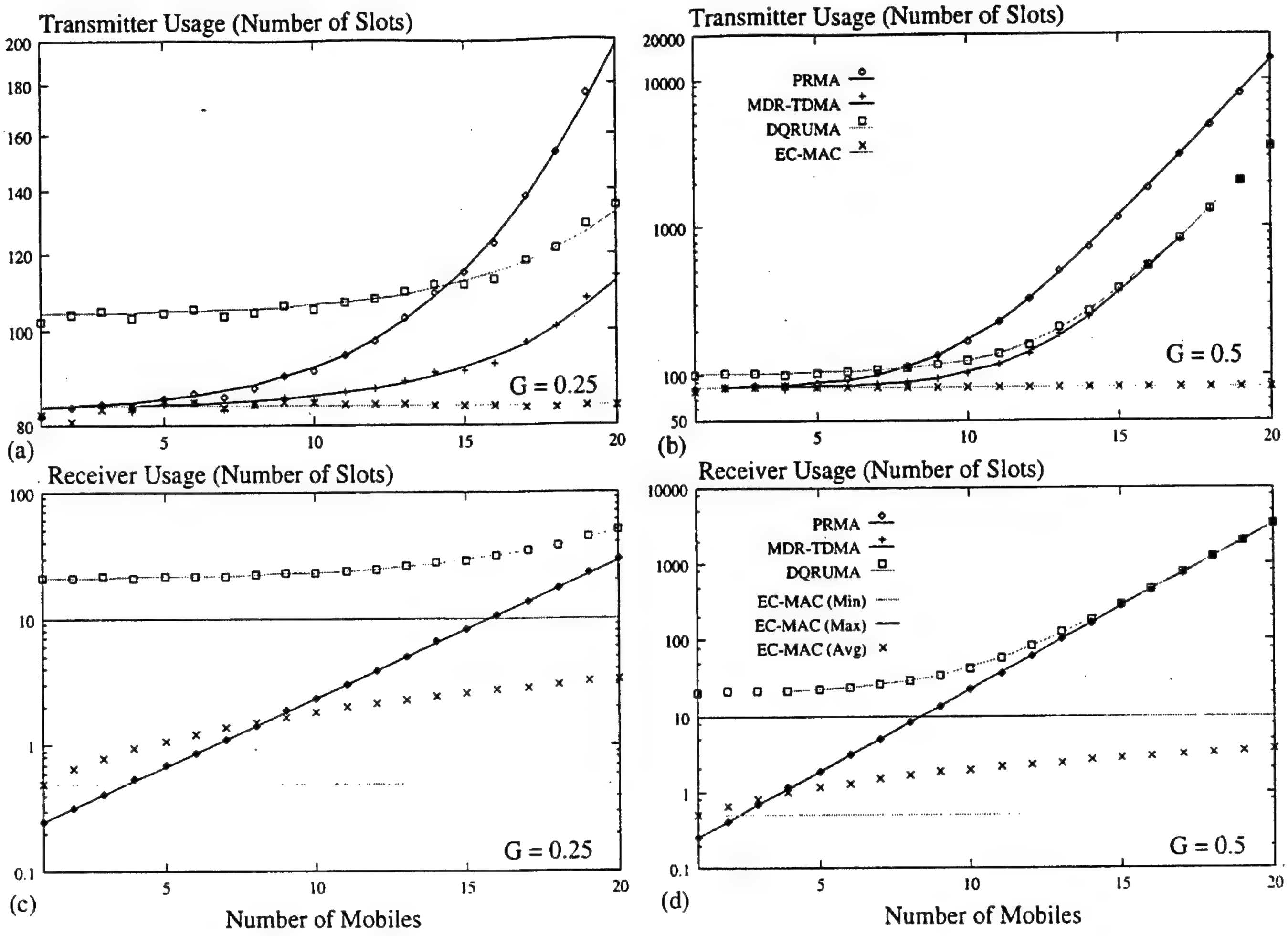


Figure 9: (a)-(b) Transmitter usage time (T_t), and (c)-(d) receiver usage time (T_r) versus number of mobiles (N) for transmitting periodic packets. The figures are plotted for $G \in \{0.25, 0.5\}$. Points represent simulation and lines represent analysis.

the mobile needs to listen to transmission permissions explicitly for every slot. Fig. 7 (c) and (d) present the results for DQRUMA when the mobile only listens to one slot for permission. Depending on traffic load and scheduling policy, the mobile may need to listen to more than one slot. Therefore, values plotted for DQRUMA represent its lower bound.

The transmitter usage time remain constant in EC-MAC in fig. 7. Fig. 7 (a) and (b) indicate that transmitter usage time is quite small in comparison to other protocols. It is very close to 802.11 when the load is heavy. Fig. 7 (c) and (d) show two lines for EC-MAC which are the minimum and maximum for the receiver to be utilized while transmitting a packet. Depending on how long the mobile listens to the transmission order and schedule beacon, the receiver usage time may be greater or less than other protocols. Simulation results presented here show the average time of all mobiles in the system. When the traffic load is light, the base station may only issue few transmission orders and schedule messages. The average receiver usage time, therefore, is closer to the lower bound. As the traffic load increased, more transmission orders and schedule

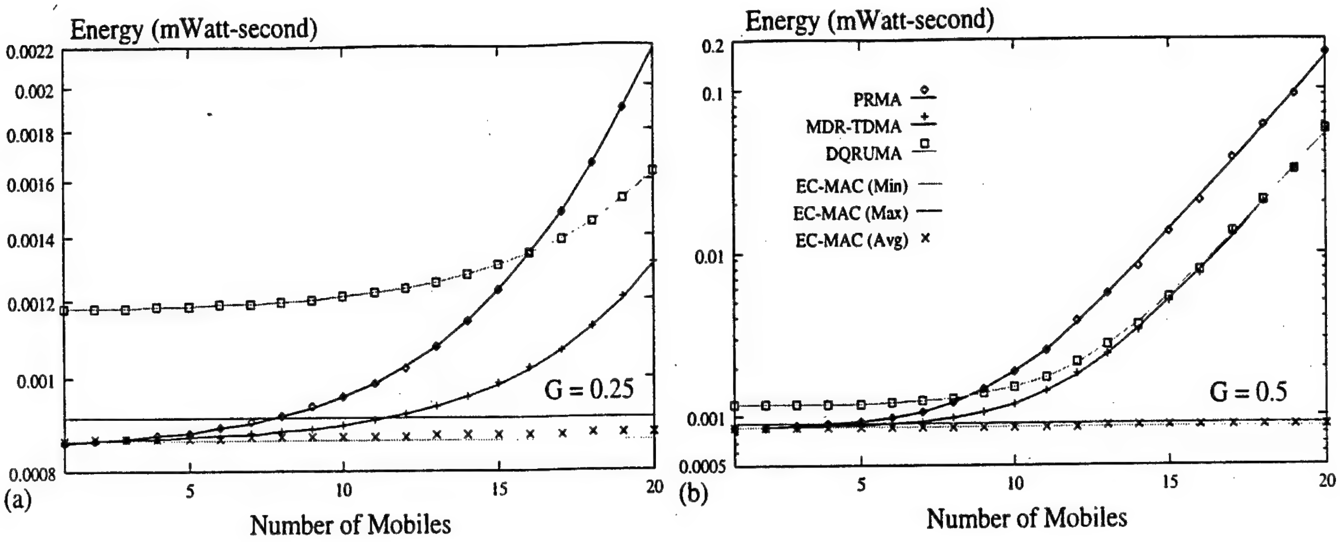


Figure 10: Energy spent per useful bit transmitted using Proxim's RangeLAN2 2.4 GHz radio, versus number of mobiles for transmitting periodic packets. The figures are plotted for $G \in \{0.25, 0.5\}$. Points represent simulation and lines represent analysis.

messages are issued. Hence, the average receiver usage time is closer to the upper bound.

Fig. 8 provides an approximate comparison of energy spent per useful bit transmitted, while transmitting a single packet using Proxim's radio card. Since MDR-TDMA and DQRUMA use the short packet for contention, they consume less energy than PRMA does. IEEE 802.11 senses the channel before transmission, so it reduces collision. However, it may need to sense several slots before it captures the medium. Therefore, 802.11 consumes more energy than PRMA, MDR-TDMA and DQRUMA do in lightly-loaded systems. On the other hand, during heavy system traffic there might be too many contentions for slotted ALOHA. We can see that 802.11 performs better than MDR-TDMA and DQRUMA when there are around 10 mobiles in fig. 8 (b). Fig. 8 also shows that the upper and lower bounds of energy consumption of EC-MAC are independent of the traffic load and number of mobiles. Simulation shows that the average energy consumption is smaller than all other protocols. In fact, we see that even the upper bound of energy consumption of the EC-MAC protocol can be significantly less than other protocols for heavily-loaded systems.

Figs. 7–8 examined transceiver utilization for a single packet. Now, Figs. 9–10 show the time usage for a voice talkspurt which is around 84 packets. PRMA, MDR-TDMA, and EC-MAC have the slots assigned for voice traffic by reservation. In DQRUMA, subsequent requests for voice packets are piggybacked on to outgoing packets. Since the voice transmission in 802.11 standard is not defined, we do not consider it in this analysis.

Fig. 9 (a) and (b) examine the transmitter and receiver usage time while transmitting a voice talkspurt. The general trends for PRMA, MDR-TDMA, and EC-MAC are similar to those for a single packet in fig. 7 (a) and (b) except that the transmitter must be powered on for all subsequent packets. In addition to voice packets, DQRUMA requires piggybacking requests for all subsequent packets as well. Hence, in lightly-loaded systems, the transmitter usage time for DQRUMA is higher than that for other protocols. We also note that DQRUMA performs better than PRMA in heavily-loaded systems. This is because PRMA transmits too many full-length packets for contention thus consuming more energy.

Fig. 9 (c) and (d) indicate that the receiver in DQRUMA needs to be turned on to receive transmission permissions for all voice packets. On the other hand, in PRMA, MDR-TDMA, and EC-MAC, the mobile has prior knowledge concerning its assigned transmission slot. In other words, it does not need to listen for permissions. The result of this difference between DQRUMA and the other protocols is that the receiver usage time in DQRUMA is higher than the others. We also see that DQRUMA performance is close to PRMA only when the offered traffic load is heavy. In fig. 9 (c) and (d), we assume DQRUMA only needs to listen to one slot for transmission permission. Depending on traffic load and scheduling policy, the mobile may need to listen to more than one slot resulting in a larger receiver usage time. Fig. 9 (c) and (d) indicates that PRMA and MDR-TDMA have the same receiver usage time. This is because we assume the length of acknowledgment in PRMA is identical to that in MDR-TDMA.

Fig. 10 presents the energy spent per useful bit transmitted while transmitting a voice talkspurt. Since the most utilized resource for periodic packets is transmitter, the energy consumption mostly depends on the transmitter usage time shown in fig. 9 (a) and (b). Figure 10 shows that EC-MAC almost consumes the least energy for periodic packets in all cases. Even the upper bound of EC-MAC is better than others when the number of mobiles is larger than 11, each with $G = 0.25$. Energy consumption in PRMA and MDR-TDMA depends the number of contentions and collisions and thus on the traffic load. MDR-TDMA uses mini-slot for contention and hence performs better than PRMA on the power perspective. Although DQRUMA uses relatively a shorter packet than PRMA does for contention, it needs more energy than PRMA does in a lightly-loaded system. This is because in DQRUMA addition burden is placed on both the receiver and transmitter to send piggybacking requests and listen to transmission permissions for all voice packets.

In general, we see that protocols should reduce the number of contentions. 802.11 senses the medium before transmitting. This results in fewer collisions than slotted ALOHA in PRMA. The receiver usage time, however, might be very large. Using short packet for contention also reduces the usage time for transmitter and receiver. In terms of energy conservation, reservation ALOHA is better than piggybacking for a message with contiguous packets. In DQRUMA, the explicit slot-by-slot announcement allows the base station to implement "optimal" and "just-in-time" scheduling. Because scheduling is done by a slot-by-slot basis, DQRUMA can potentially reduce packet latency. However, the additional burden placed on the receiver sub-system to receive and decode during every slot tends to increase energy consumption.

6 Summary

This paper considers mobile battery power conservation from the media access layer protocol perspectives in wireless networks. Energy conservation has typically been considered at physical layer issues, and to a certain extent at the access protocol level. The paper describes various energy conservation techniques proposed in different access protocols including IEEE 802.11, PRMA, MDR-TDMA, EC-MAC, and DQRUMA. The observations from the analysis and a qualitative comparison of the different protocols are presented. The analysis here shows that protocols that aim to reduce the number of contentions perform better from an energy consumption perspective. The receiver usage time, however, tends to be higher for protocols that require the mobile to sense the medium before attempting transmission. For messages with contiguous packets, our analysis shows that reservation is more energy conservative than piggybacking. For example, IEEE 802.11 minimizes the transmitter usage time while sending a packet, but requires the receiver to turned on for the longest period of time for channel sensing. On the other hand, EC-MAC uses an explicit transmission order for sending reservations and a broadcast schedule to reduce energy consumption.

References

[1] M. Naghshineh (Guest Ed.), "Special issue on Wireless ATM," *IEEE Personal Communications*, vol. 3, Aug. 1996.

[2] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. B. Srivastava, and J. A. Trotter, "SWAN: A mobile multimedia wireless network," *IEEE Personal Communications*, vol. 3, pp. 18–33, Apr. 1996.

[3] K. M. Sivalingam, M. B. Srivastava, P. Agrawal, and J.-C. Chen, "Low-power access protocols based on scheduling for wireless and mobile ATM networks," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, (San Diego, CA), pp. 429–433, Oct. 1997.

[4] D. Raychaudhuri and N. D. Wilson, "ATM-based transport architecture for multi-services wireless personal communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 1401–1414, Oct. 1994.

[5] D. Raychaudhuri, L. J. French, R. J. Siracusa, S. K. Biswas, R. Yuan, P. Narasimhan, and C. A. Johnston, "WATMnet: A prototype wireless ATM system for multimedia personal communication," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 83–95, Jan. 1997.

[6] M. J. Karol, Z. Liu, and K. Y. Eng, "An efficient demand-assignment multiple access protocol for wireless packet (ATM) networks," *ACM/Baltzer Wireless Networks*, vol. 1, no. 3, pp. 267–279, 1995.

[7] J.-C. Chen, K. M. Sivalingam, and R. Acharya, "Comparative analysis of wireless ATM channel access protocols supporting multimedia traffic," *ACM/Baltzer Mobile Networks and Applications*, 1998. To appear.

[8] IEEE, "Wireless LAN medium access control (MAC) and physical layer (PHY) Spec." P802.11/D5, Draft Standard IEEE 802.11, May 1996.

[9] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, and B. Ramamurthi, "Packet reservation multiple access for local wireless communications," *IEEE Transactions on Communications*, vol. 37, pp. 885–890, Aug. 1989.

[10] K. M. Sivalingam, M. B. Srivastava, and P. Agrawal, "Low power link and access protocols for wireless multimedia networks," in *Proc. IEEE Vehicular Technology Conference*, (Phoenix, AZ), pp. 1331–1335, May 1997.

[11] K. Govil, E. Chan, and H. Wasserman, "Comparing algorithms for dynamic speed-setting of a low-power CPU," in *Proc. ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, (Berkeley, CA), pp. 13–25, Oct. 1995.

[12] B. Marsh, F. Douglis, and P. Krishnan, "Flash memory file caching for mobile computers," Tech. Rep. MITL-TR-59-93, Matsushita Info Tech Lab, Princeton, NJ, June 1993.

[13] F. Douglis, P. Krishnan, and B. Marsh, "Thwarting the power-hungry disk," in *Proc. Winter 1994 USENIX Conf*, pp. 293–306, Jan. 1994.

[14] B. M. Gordon, E. Tsern, and T. H. Meng, "Design of a low power video decompression chip set for portable applications," *Journal of VLSI Signal Processing Systems*, vol. 13, pp. 125–142, 1996.

[15] A. Fox, E. A. Brewer, S. Gribble, and E. Amir, "Adapting to network and client variability via on-demand dynamic transcoding," in *Proc. ASPLOS-VII*, Oct. 1996.

[16] R. Alonso and S. Ganguly, "Energy efficient query optimization," Tech. Rep. MITL-TR-33-92, Matsushita Info Tech Lab, Princeton, NJ, Nov. 1993.

[17] T. Imilienski, S. Vishwanathan, and B. R. Badrinath, "Energy efficient indexing on air," in *Proc. ACM SIGMOD*, May 1994.

[18] M. Stemm and R. H. Katz, "Measuring and reducing energy consumption of network interfaces in hand-held devices," *IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Science*, Aug. 1997.

[19] M. Zorzi and R. R. Rao, "Error control and energy consumption in communications for nomadic computing," *IEEE Transactions on Computers*, vol. 46, pp. 279–289, Mar. 1997.

[20] M. Zorzi and R. Rao, "Energy constrained error control for wireless channels," *IEEE Personal Communications*, Dec. 1997.

[21] P. Lettieri, C. Fragouli, and M. B. Srivastava, "Low power error control for wireless links," in *Proc. ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, (Budapest, Hungary), Sept. 1997.

[22] J. M. Rulnick and N. Bambos, "Mobile power management for wireless communication networks," *ACM/Baltzer Wireless Networks*, vol. 3, pp. 3–14, Mar. 1997.

[23] Proxim Inc., "Proxim: the leader in wireless LANs." http://www.proxim.com/, 1998.

[24] Lucent Technologies, "The age of wireless LANs has arrived." http://www.wavelan.com/, 1996.

[25] D. Petras and A. Kramling, "MAC protocol with polling and fast collision resolution for an ATM air interface," in *Proc. IEEE ATM Workshop*, (San Francisco, CA), Aug. 1996.

[26] K. S. Natarajan, "A hybrid medium access control protocol for wireless LANs," in *Proc. 1992 IEEE International Conference on Selected Topics in Wireless Communications*, (Vancouver, B.C., Canada), June 1992.

[27] ETSI-RES10, "High performance radio local area network (HIPERLAN)." ETS 300, Feb. 1997.

[28] G. L. Stüber, *Principles of Mobile Communication*. Kluwer Academic Publishers, 1997.

[29] J. Walrand, *Communication Networks*. Aksen Associates, Inc., 1991.

[30] L. Kleinrock and F. A. Tobagi, "Packet switching in radio channels: Part I - Carrier sense multiple-access modes and their throughput-delay characteristics," *IEEE Transactions on Communications*, vol. COM-23, pp. 1400–1416, Dec. 1975.

[31] R. Rom and M. Sidi, *Multiple Access Protocols: Performance and Analysis*. Springer-Verlag, Inc., 1990.

[32] D. J. Goodman and S. X. Wei, "Efficiency of packet reservation multiple access," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 170–176, Feb. 1991.

[33] J. R. Jump, *YACSIM Reference Manual*. Rice University, Department of Electrical and Computer Engineering, 2.1 ed., Mar. 1993.

[34] A. M. Law and W. D. Kelton, *Simulation Modeling and Analysis*. McGraw-Hill, Inc., 1991.

SUBMITTED TO ACM/IEEE TRANS. ON NETWORKING, 1997.

Battery Power Level based Adaptive Power Control and Scheduling Algorithms for CDMA Wireless Networks

Shalinee Kishore[1], Jyh-Cheng Chen[2], Krishna M. Sivalingam[3], and Prathima Agrawal[4] *

[1]WINLAB, Rutgers University, Piscataway, NJ 08855

[2]Department of Electrical & Computer Engineering, State University of New York at Buffalo, Buffalo, NY 14260

[3]School of Electrical Engineering & Computer Science, Washington State University, Pullman, WA 99164

[4]Networked Computing Technology Department, AT&T Labs, Whippany, NJ 07981

Abstract

Energy efficiency of portable terminals is an important issue in mobile wireless networks. A general constraint common to many wireless networks lies in the short lifetime of mobile terminal batteries. Energy efficient protocols that adapt to terminal battery power level can be used to lessen this limitation. This paper addresses such power efficient adaptive algorithms in medium access control (MAC) protocols and in power control mechanisms for hybrid CDMA/TDMA wireless networks. Our access protocol dynamically schedules CDMA channels to mobiles based on their traffic requests and battery power levels. One technique clusters low-power mobiles into one or more slots and adjusts the power control algorithm such that these mobiles can transmit with desired reliability, i.e. target transmission error rate, despite the low-power conditions. Another technique adjusts the power control algorithm parameters such as the amount of power increments/decrements, the periodicity of power control updates, the transmit power levels, and the number of simultaneous transmitters. Discrete-event simulation has been used to demonstrate that the proposed techniques provide low-power mobiles with increased throughput and reduced latency. This analysis indicates that low-power users not only improve their throughput and delay performance but also transmit information using less battery power. Also, using our algorithm low-power terminals transmit information with target error rates and thus help maintain quality-of-service (QoS). The impact of our scheduling algorithm on the transmission at mobiles with high battery power is also studied.

Keywords: Code Division Multiple Access (CDMA), Scheduled CDMA, Power control, Low-power operation, Wireless networks, Mobile networks.

1 Introduction

The rapid penetration of wireless services like cellular voice, Personal Communication Services (PCS), mobile data, and wireless LANs and PBXs in recent years is an indication that users place significant value on

*CONTACT: Krishna Sivalingam, School of Electrical Engg. and Computer Science, Washington State University, Pullman, WA 99164. Phone: (509) 335-3220, Fax: (509) 335-3818. Email: krishna@eecs.wsu.edu. The authors can be reached at kishore@ece.rutgers.edu, jcchen@eng.buffalo.edu, krishna@eecs.wsu.edu, and pa@research.att.com.

portability as a key feature in their telecommunication needs [1]. The next generation of wireless systems aims to extend these current services to include multimedia applications. Driven by this forecast of multimedia integration and by the users' growing dependency on portability, the primary focus of recent wireless networking research has been to develop architectures to accommodate seamless communications while ensuring quality-of-service (QoS) transmission of multimedia traffic [2–4].

An implication of terminal portability is the use of batteries as power supply for mobile terminals. Since batteries provide limited energy, a general constraint on wireless communications lies in the battery lifetime. Due to this limitation, it has been proposed that low-power design should also be a crucial consideration in designing all layers of the protocol stack for wireless networks [5]. Low-power design at the hardware layers uses different techniques including variable clock speed CPUs, flash memory, disk spindowns [5]. At the application layer, low-power video compression, transcoding at the basestation and energy efficient database operation have been considered [6]. In [7], the power drained by the network interface in hand-held devices was studied. A power efficient probing scheme for error control in link layer is proposed in [8]. In addition to hardware considerations, the wireless infrastructure should use information about user's battery level and adapt network operation accordingly.

As discussed in [5], the CPU, the transmitter, and the receiver are the major consumers of battery power at the mobile terminal for access protocol activities. The integration of multimedia traffic only adds to the processing requirements and in turn increases power usage at the mobile station. To reduce the burden placed on low-power users, we propose modifications to the medium access and data link layer of the protocol stack and to the power control mechanism for hybrid CDMA/TDMA wireless network systems. More specifically, our adjustments are geared at the scheduler that assigns transmission times to mobiles in a particular coverage area. In our proposed approach, mobiles periodically transmit their current battery status to their serving basestations (BS). Once aware of mobile battery levels, the network architecture operating under our scheduling scheme can then increase the total throughput of low-power users while reducing their latency. The performance of low-power mobiles can be further enhanced using a power control process that allows for energy-conserving transmission.

The medium access control (MAC) protocol presented here is derived from the Energy-Conserving Medium Access Protocol (EC-MAC) described in [4]. EC-MAC relies – for energy conservation reasons – on scheduling algorithms to assign transmission times to mobiles. In this paper we develop a scheduling algorithm explicitly suited to benefit low-power (LP) users in a hybrid CDMA/TDMA multimedia network. At the beginning of every time frame, users transmit their traffic needs – traffic type and queue status – to the BS. The BS processes the requests of all the users and incorporates their battery power information to generate an uplink schedule that it then broadcasts to the mobiles. The proposed algorithm prioritizes LP mobile traffic to be scheduled for transmission before high-power (HP) mobile traffic. At the same time, the scheduler also considers the registered priority of the traffic – for example, real-time and non real-time.

One ramification of this scheduling algorithm is that LP users are assigned adjacent transmission slots and hence are "clustered" in time. The modifications proposed here are a result of this "clustering" phenomena and affect the power control algorithm. All systems, particularly CDMA-based ones, use power control algorithms to govern the transmission power levels of the mobiles [9]. Additionally, power control is necessary to combat the near-far effect and to increase CDMA system capacity [10]. Since mobiles consume most power in the transmit mode [5], the BS assigns lower transmit power levels to those mobiles scheduled during LP "clusters". To perform this effectively, the BS has to employ an adaptive power control algorithm that adjusts the necessary parameters to achieve desirable transmit powers for each "cluster".

Power control algorithms for CDMA-based systems also employ a closed-loop power control mechanism by which the BS periodically informs each mobile to increase or decrease its transmit power to meet a prescribed signal-to-interference ratio (SIR) [9, 11]. Based on the presence of LP users, we propose that the BS can adjust these power control updates to occur less frequently at LP mobiles. In addition to altering the frequency of the updates, the closed-loop power control mechanism can ask mobiles to increase or decrease their transmit powers in variable increments/decrements based on their battery status. One implication of these modifications to the closed-loop power control lies in an increased error rate. To combat this, we exploit the time-division properties of the hybrid system. This can be done using a variable number of simultaneous transmitters and a target SIR that depends on the current "cluster."

The performance of a simple single-cell system operating under our proposed modifications was studied using discrete event simulation. Two different schedulers and power control techniques were implemented: one where only traffic priority was used for scheduling and the other where both traffic and battery level was used by the scheduler and by the power control algorithm. Comparisons between the throughput and average packet delay of the two systems are presented. These results indicate that due to their higher scheduling priority, LP mobiles significantly improve their throughput and delay performance when operating under battery power adaptation. We also compare the two systems in terms of their power efficiency by computing the total power consumed per transmitted packet. Once again, due to adaptive measures taken by the scheduler and the power control algorithm, LP users were able to consume less battery power during transmission. Furthermore, using the comparisons, we illustrate that this energy efficiency is gained without lessening QoS requirements, i.e. by maintaining the target transmission error rate.

The rest of the paper is organized as follows. Section 2 provides the background on CDMA/TDMA access and power control. Section 3 provides the details of the mobile queuing architecture and the access protocol. Section 4 describes the techniques studied in this paper to support transmission from LP mobiles. Finally simulation results are collected and presented in Section 5. Section 6 summarizes the paper.

2 Background on CDMA/TDMA

This section provides background material on CDMA/TDMA and power control.

System Description: The paper considers an infrastructure-based wireless network, with a BS serving a region called the *cell* where a set of mobiles in this cell are served by the BS. A set of code division multiple access (CDMA) channels is available in the cell for communication. CDMA channels can either be a sequence of carrier frequencies for *frequency-hopped* CDMA or a sequence of binary symbols as in *direct sequence* CDMA [1,12–14]. The BS has the responsibility of coordinating mobile access to the channels in the current cell.

Some of the available CDMA channels are used for downlink (BS-to-mobile) transmission, and the other channels are used for uplink (mobile-to-BS) transmission. Another way to separate uplink and downlink channels is through the use of frequency division duplexing (FDD) where two separate frequency bandwidths are assigned for either BS-to-mobile or mobile-to-BS communications.

For the hybrid CDMA/TDMA system under consideration here, time is divided into equal-length *slots* on each of the channels. Each uplink channel is allocated to one of the mobiles for a specified number of slots. During its assigned slot(s), the mobile must power up its transmitter to a specified transmit power level and – using the predetermined CDMA code as channel – send out its digital stream which maybe buffered. When not transmitting information, i.e. when a mobile is not assigned to the current slot, the terminal still communicates with the base at a lower power level for synchronization purposes. A two-dimensional array of CDMA codes and time slots is defined and is considered in detail in the next section. Each CDMA code can be allocated to exactly one mobile in a time slot. Thus, multiple parallel communication channels are established within the cell using different pseudo-random codes. In our analysis we will deal with CDMA codes consisting of binary sequences, i.e. DS-CDMA.

CDMA Power Control: Mobiles operating in CDMA-based systems transmit under strict power control. One of the implications of spreading a message signal over a wideband is that each transmitted signal must be received by the BS at similar power levels in order to maximize the total user capacity. Thus mobiles that are located farther away from the BS must transmit their signals at a much higher power level than mobiles positioned near a BS so that both signals may be received at the BS at equivalent power levels. This requirement needed to combat the *near-far problem* is of great importance in CDMA. As described in [9,11], CDMA systems employ a power control algorithm consisting of open-loop and closed-loop power control at the mobile as a means to counter the near-far effect.

In open-loop power control, each mobile measures the signal power level of the downlink message it received from the BS. Based on this measurement and a prescribed target, the mobile then computes how much to adjust its own transmission to achieve the desired target power signal or message level. In other words, the channel loss disparity is adjusted individually at each mobile by fixing its transmit power based on the measured received power [9]. This disparity is the open-loop power control, P_{Open}, and can be represented as [14]:

$$P_{Open} = P_{Target} - P_{Received} \tag{1}$$

where P_{Target} is the target power level for the particular BS and $P_{Received}$ is of course the received power measurement. Note that the exact value of P_{Target} depends on the propagation characteristics in the current coverage area and is therefore basestation specific. This target level is transmitted to each mobile in the current cell as a part of the overhead information.

For closed-loop power control, the BS receives each mobile's signal and measures its power level. Using this measurement, the BS determines if each received power level suffices an exact target value. Then based on this decision, the BS periodically multiplexes a power control message in the downlink data directed to each of the transmitting mobiles. This message indicates to the mobile if it should increase or decrease its transmitted power so as to maintain the desired power levels at the BS. The amount of power increase or decrease, $\pm\Delta$, is prescribed by the system. For example, a BS operating under Interim-Standard 95 multiplexes a one bit power control message every 1.25 ms where a bit "one" indicates that the mobile should increase its power by 1 dB and a "zero" indicates it should decrease its power by 1 dB [15].

The closed-loop power control adjustment, P_{Closed}, therefore, is:

$$P_{Closed} = \pm\Delta \tag{2}$$

The final power adjustment performed at the mobile depends on a combination of these open and closed-loop power control algorithms. The mobile computes how much to increase or decrease its power level based on the open-loop measurement and then listens to the BS to determine the closed-loop adjustment. The mobile adds the two open and closed-loop adjustments to compute its final amplification factor. This entire control process is commonly referred to as the "bang-bang" control loop [9]. The adjusted transmit power, $P_{NewTransmit}$, is then

$$P_{NewTransmit} = P_{Transmit} + P_{Open} + P_{Closed} \tag{3}$$

In addition to the "bang-bang" control loop, the power control mechanism in hybrid CDMA/TDMA systems is also responsible for assigning transmit power levels to each mobile. Consider a single-cell DS-CDMA/TDMA system with a system bandwidth of W. Those mobiles that do not transmit during a particular slot synchronize with the BS at power P_o. For our current analysis we assume that standby power is constant for all other users and low enough to ignore. Each mobile, i, assigned to the current slot is given a transmit power P_i by the BS. The energy required at terminal i to transmit one bit, E_b is merely GP_i, where G represents the spreading factor or the length of the binary CDMA code. So to transmit one information bit, the terminal has to transmit a sequence of G bits or chips each at power P_i. While terminal i attempts to transmit its bit, the total interference, I_o, in the system is the sum of the transmit powers of all other transmitting mobiles plus the total Gaussian noise in the system. So, the ratio $(\frac{E_b}{I_o})_i$ which also represents the SIR for mobile i is given by [10]:

$$\left(\frac{E_b}{I_o}\right)_i = \frac{GP_i}{\sum_{\{j=1,\, j\neq i\}}^{N} P_j + \eta_o W} \tag{4}$$

where η_o is power spectral density of the additive white Gaussian noise present in the system and N is the number of mobiles transmitting during the current slot. Assume that each mobile has the same minimum SIR requirement, γ, for the transmitted information. This γ is part of the QoS constraint for the particular traffic type of the transmitted stream. The assumption here, therefore, is that all the mobiles in the current slot are transmitting the same information type. This QoS constraint on the SIR can be represented as:

$$\frac{GP_i}{\sum_{\{i=1,\, j\neq i\}}^{N} P_j + \eta_o W} \geq \gamma \tag{5}$$

The task of assigning transmit power levels in hybrid CDMA/TDMA systems is to allocate the minimum total transmitted power while meeting the QoS constraint above. As shown in [10], the optimal solution to this allocation, if it exists, is met when:

$$\frac{GP_i}{\sum_{\{j=1, j\neq i\}}^{N} P_j + \eta_o W} = \gamma \tag{6}$$

Following the procedures illustrated in [10, 16], we can solve for the optimal P_i as:

$$P_i = \frac{\eta_o W g}{1 - Ng} \tag{7}$$

where

$$g = \frac{\gamma}{\gamma + G} \tag{8}$$

The existence of this optimal solution depends on the constraints on P_i. If it is merely required that all transmit power levels P_i must be positive, then the optimal solution to the constraint in (6) is met when [10]:

$$Ng < 1 \tag{9}$$

The standard power control mechanism in CDMA/TDMA hybrid systems is therefore also responsible for assigning the optimal power levels for each mobile in each slot based on the SIR constraint and transmit power requirement. These SIR constraints are used to ensure a transmission error rate for the information stream and subsequently are used to ensure QoS.

3 Architecture and Access protocol

This section describes the mobile architecture including the queuing structure and the hybrid access protocol.

3.1 Mobile architecture

A mobile has a transmitter and a receiver, each capable of transmitting and receiving signals using a subset of, and possibly all, the available channels in the current cell. The mobile battery has a limited lifetime, and one of the main objectives of this paper is to conserve battery power usage with a two part process. First, medium access control (MAC) related activities, including data transmission and reception, are restructured so as to realize and appropriately react to a mobile's battery power level and transmission queue. Second, as a natural consequence to this modification of the MAC activities, the power control algorithm inherent in CDMA systems is altered to adapt to mobiles' battery power levels.

A mobile can originate and terminate multiple data connections, that enable it to communicate with other computers and communication devices. All communication to and from the mobile is through the BS. Each such connection is referred to as a Virtual Circuit (VC). This technique is adopted in ATM (Asynchronous Transfer Mode) networking for multimedia communication [17]. Each VC is associated with a transmission priority established by the mobile application utilizing this VC for communication. These priorities will be utilized by the BS when allocating channels to the mobiles. Each mobile maintains a separate queue for each of its VC, as shown in Figure 1. Information arrives at each queue in the form of a *packet* and is buffered until transmission.

3.2 Hybrid Access Protocol

A hybrid medium access protocol that combines CDMA and TDMA is studied in this paper. The access protocol is derived from the EC-MAC protocol defined in [4].

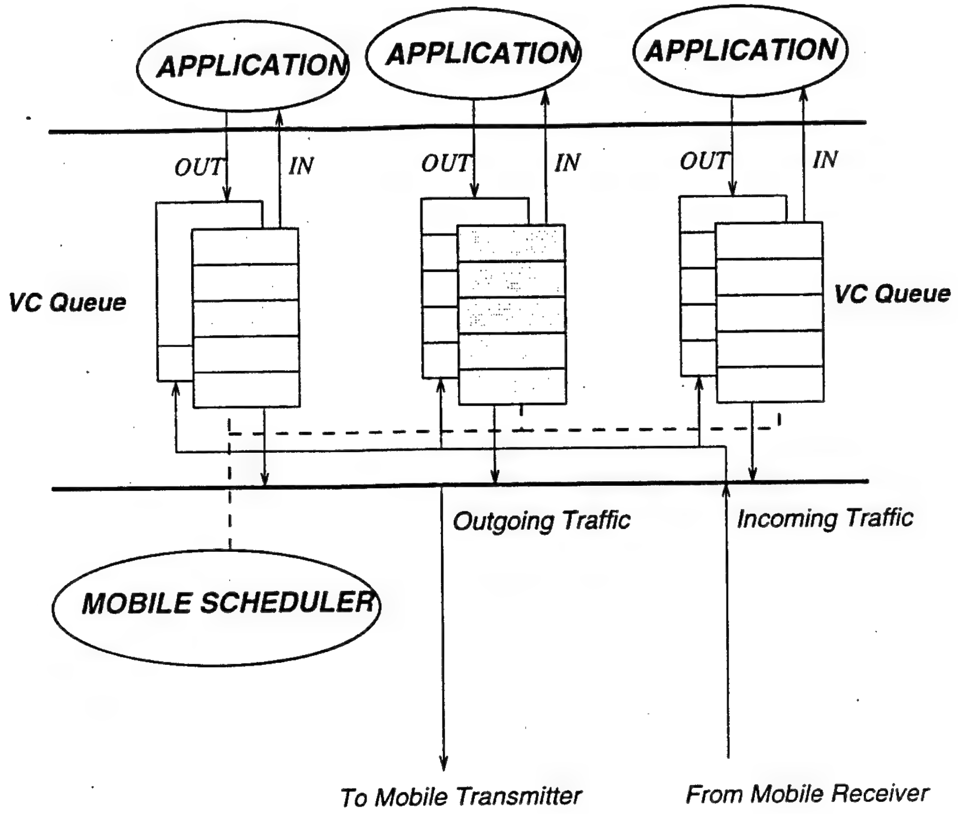


Figure 1: Virtual Circuit Queue structure within a mobile.

Transmission in the network is organized into *frames* which is further divided into subframes. Figure 2 shows a frame with three subframes. The following is a description of the frame structure.

1. At the beginning of each frame, there is a frame synchronization phase that aids new and current users to establish and maintain synchronization.

2. In the *request/update and new user* phase, mobiles use two distinct, known sets of uplink and downlink channels. Using the *request/update* set of channels, registered mobiles transmit their current queue status, battery power level, and other information on the uplink channels to the BS. During this same time, new mobiles entering the system register at the BS using *new-user* set of channels.

3. Next comes the *downlink broadcast* phase, when the BS broadcasts data, acknowledgments, and scheduling information that all mobiles need to receive. In CDMA, this downlink data also includes power control information for the mobile. The power control information passed on to the mobiles includes closed-loop power control updates as well as assigned transmit power levels.

4. The *downlink unicast/multicast and uplink* phase follows. During this time two distinct set of channels are used once again. During this phase a group of downlink channels are used so the BS may transmit unicast or multicast data on different channels. Since CDMA mobiles are capable of receiving

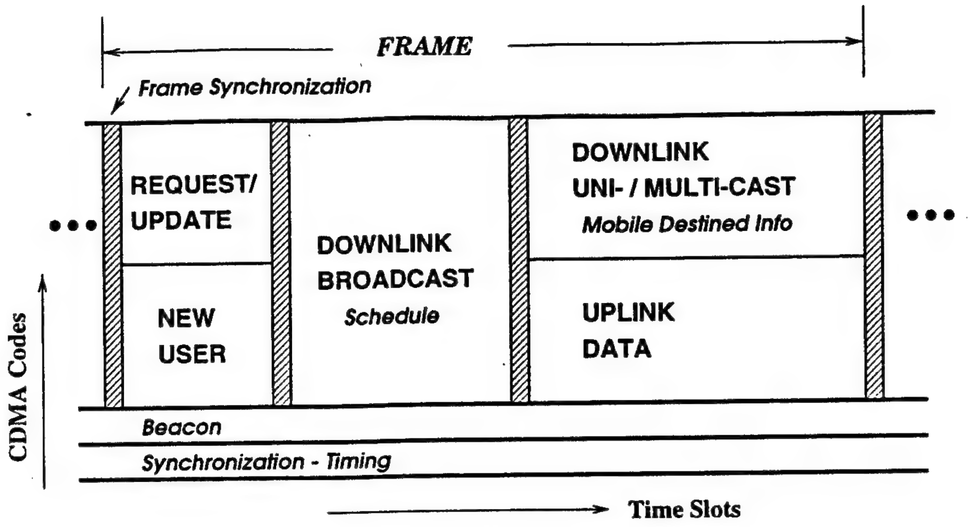


Figure 2: Frame structure of the multiple access protocol showing the various sub-phases.

and transmitting simultaneously, during this phase mobiles can also transmit data on the uplink using their assigned CDMA codes and time slots. Once again closed-loop power control messages can be multiplexed into the downlink streams.

In addition to these information channels, there are two additional downlink channels: *beacon* and *synchronization and timing*. On these channels, the BS continuously transmits overhead information. On the *beacon*, the BS informs mobiles what channels are allocated for the *request/update* phase and for the *new user* phase. It carries additional information concerning phasing and transmit signal power levels. The *synchronization-timing* channel provides system timing information.

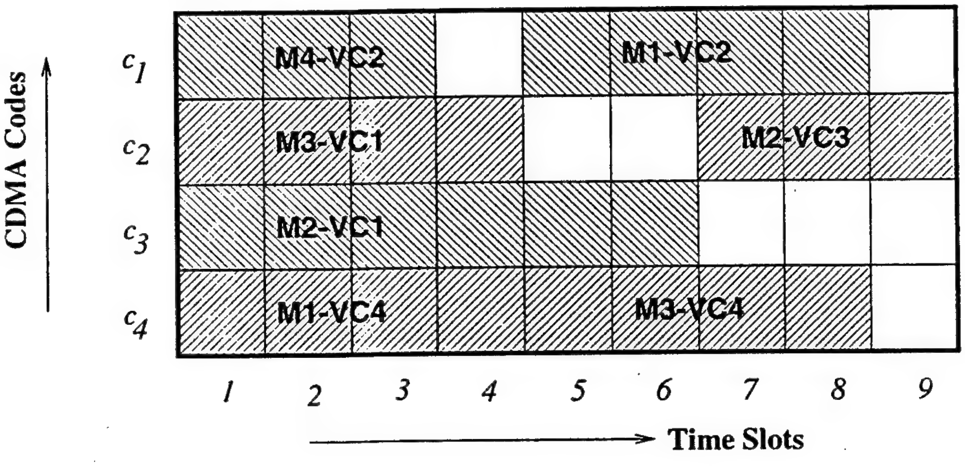


Figure 3: Sample allocation of CDMA code/time-slot combination to different mobiles. Each allocation is for a specific Virtual Circuit associated with a mobile.

Periodically, each active mobile in the current cell transmits the current queue status of all its VC queues.

The BS executes a scheduling algorithm which generates an allocation of channels and time slots to the mobiles. The order of this allocation is calculated based on the priorities associated with the different VCs. Specifically, the BS assigns a code-time slot combination to a mobile's VC, an example of which is shown in Figure 3. The allocation information is then broadcast to all the mobiles which transmit according to this schedule.

4 Proposed Algorithms

The techniques proposed to adapt network operations to lower-power terminals are described in this section. The algorithm operates under the assumption that each BS has knowledge of the battery power level of all mobiles in its coverage area. Thus the first requirement of our algorithm is a periodic battery power level update when mobiles transmit their current battery supply to their serving BS. This update occurs during the *request/update and new user* phase of the transmission frame. Based on a simple threshold comparison, the BS groups mobiles of similar power levels together. Our description here and the system simulation assumes that mobiles are classified into two types: HP and LP (high-power and low-power). The algorithm however can easily be generalized to more number of discrete power levels. The VC queues at each mobile are also classified based on their QoS priority. Here we assume two priority levels – high-priority and low-priority. Again, the algorithm can be easily generalized to more priority levels.

4.1 Scheduling Adjustments

Once mobiles have been characterized according to battery level, the scheduling algorithm allocates slots in the *uplink data phase* of the MAC frame to each VC queue based on terminal battery power level and traffic demand. One of the aims of our algorithm is to reduce the latency at the LP mobile VC queues. To do this, we must schedule transmission of packets from these LP mobiles as early as possible in the *uplink* phase. In other words, we must prioritize transmission of LP VC queues before HP VC queues.

An added dimension to this scheduling lies in the prescribed priority of each VC queue. As mentioned earlier, VC queues are placed in priority classifications independent of their battery power levels. The higher priority classifications maintain their own minimum delay and error rate requirements. These specifications are used to quantify the QoS of that VC. In order to meet these delay requirements and thus help guarantee QoS for all users in the current cell, the scheduling algorithm at the BS must designate earliest possible slots to high-priority VC queues regardless of battery power level.

These two considerations of battery power levels and VC queue priority results in the following general scheduling scheme. The BS breaks down the *uplink* phase into four intervals or "clusters." The "clusters" are defined by the combination of the two scheduling parameters: LP & high-priority, HP & high-priority, LP

& low-priority, and finally HP & low-priority. If VC queue priority was the only scheduling consideration, then BSs would allocate the first available slots to high-priority queues and then the remaining slots would be assigned to low-priority queues based on their minimum delay QoS criterion. With the knowledge of mobile battery status, we propose that the high-priority slots should first be assigned to LP users thus forming a LP & high-priority cluster at the start of the *uplink* phase. The HP & high-priority mobiles will then be given the next cluster, followed by LP & low-priority and then the HP & low-priority.

4.2 Power Control Adjustments

As a consequence of our scheduling adjustments, we observe that low power users transmit during two specific time intervals of the *uplink* phase. Since the BS handles the scheduling of the uplink slots, it has exact knowledge about the start and end times of these intervals. Due to this knowledge, we now introduce a power control mechanism at the BS that dynamically adjusts its power specifications to adapt to the presence of *all* LP users during a particular cluster. The BS knows that transmission from LP mobiles constitutes a particular segment of the *uplink data phase*. It can then reduce its power control requirements during that segment of time. Power control adjustments can, therefore, be done at the cluster level to aid LP users.

Closed-Loop Control: The first modification for the power control algorithm lies in the closed-loop requirements. As mentioned previously, the BS periodically informs users to increase or decrease their transmit power levels by a system-defined level, ($\pm\Delta$). The period for this update is also a system-defined parameter. To accommodate the low battery status of LP users, we propose that the closed-loop update requirements be lessened in any or all of the following ways:

1. First, if the mobile under consideration has been marked as LP, then the BS can make transmission power increase requests of the mobile if the power level falls below some minimum threshold. This bare minimum threshold corresponds to a value close to the call dropping SIR level and is therefore below the target SIR level. Thus, power increase requests can occur with a larger periodicity or only when absolutely required.

2. In the case of a LP mobile that is currently transmitting at a power level higher than the desired target, the BS can proceed as before and periodically multiplex a power control message that requests a transmit power decrease. Decrements of transmit power levels are favorable for battery power conservation. Therefore, a power decrease message geared towards these LP mobiles can be dynamically adjusted to occur as before or with greater frequency in every frame.

3. The BS can adapt its power control scheme to require power decrease in larger possible decrements so that fewer power decrease control messages have to be sent to the mobile before the mobile can

achieve its target power level. The power control algorithm can ask LP mobiles for power increases in smaller increments so as to achieve the bare-minimum power level while expending the least amount of energy possible.

4. The final and perhaps the most extreme alternative is if the BS decides to completely backoff from any power increase requests realizing the low battery status of these mobiles. So in effect, the BS avoids asking a LP mobile to expend its remaining power to meet a higher closed-loop power control requirement. In this situation, the BS relies on the dynamic nature of the radio channel, the open-loop power control, as well as the robustness of CDMA error-correcting codes to correctly receive the transmitted signal.

As discussed in [18], the fading characteristics of the wireless link leads to imperfect power control, i.e. a disparity between the desired and measured SIR. The modifications proposed above perpetuate this disparity. One way to combat the deterioration brought upon by imperfect power control is to increase the required SIR [9]. Since our modifications apply only to those mobiles transmitting during LP clusters, we note that for reliable transmission we must increase our basic SIR requirements during this interval. The ramifications of increasing the target SIR is discussed later.

Transmit Power Assignment: The next adaptation of the power control lies in the assignment of transmit power levels. LP mobiles can conserve their battery supply if they transmit at lower power levels. This can be achieved by constraining the transmit power levels of LP users during their cluster. Let P_{LPMax} represent the maximum power level that a LP mobile can transmit at during its cluster. Note that if P_{Max} is the maximum power at which all mobiles transmitted before battery power adaptation, then due to power conservation for LP mobiles, we require that $P_{LPMax} < P_{Max}$. Additionally, note that the HP users can still transmit at P_{Max} during HP clusters since they are not under energy constraints.

Our adjusted power allocation is similar to equation (6) with the additional pair of constraints:

$$0 < P_i \leq P_{LPMax} \quad \text{for LP mobiles} \tag{10}$$

$$0 < P_j \leq P_{Max} \quad \text{for HP mobiles} \tag{11}$$

Using equation (6) and the above transmit power limitation, the constraint in (9) can be written as:

$$N_l g_l \leq 1 - \frac{g_l \eta_o W}{P_{LPMax}} \quad \text{during LP slot} \tag{12}$$

$$N_h g_h \leq 1 - \frac{g_h \eta_o W}{P_{Max}} \quad \text{during HP slot} \tag{13}$$

where N_l is the number of LP users in a slot during a LP cluster and N_h is similarly the number of HP users in a HP slot. Assuming that γ_l and γ_h are the target SIRs and G_l and G_h are the processing gains for the LP and HP slots respectively, the parameters g_l and g_h can be defined as the following:

$$g_l = \frac{\gamma_l}{\gamma_l + G_l} \tag{14}$$

$$g_h = \frac{\gamma_h}{\gamma_h + G_h} \tag{15}$$

Note that equation (6) was presented as the power allocation requirement *for each transmission slot.* In our case, we have extended the allocation constraint to deal with transmission *during each cluster.*

Also note that since each cluster has been defined by the traffic type, i.e. either low-priority or high-priority, all mobiles in a particular cluster have the same SIR requirement. We therefore have four different SIR requirements, one for each cluster: γ_{lh} for LP & high-priority, γ_{ll} for LP & low-priority, γ_{hh} for HP & high-priority, and finally γ_{hl} for HP & low-priority. To meet the SIR requirements, each cluster also has its own processing gain, G_{lh}, G_{ll}, G_{hh}, and G_{hl}. For our analysis, let us assume that the high-priority traffic has a higher SIR requirement than the low-priority. So the allocation requirement is one of the following, depending on the cluster:

$$N_{lh} g_{lh} \leq 1 - \frac{g_{lh} \eta_o W}{P_{LPMax}} \quad \text{during LP \& high-priority cluster} \tag{16}$$

$$N_{ll} g_{ll} \leq 1 - \frac{g_{ll} \eta_o W}{P_{LPMax}} \quad \text{during LP \& low-priority cluster} \tag{17}$$

$$N_{hh} g_{hh} \leq 1 - \frac{g_{hh} \eta_o W}{P_{Max}} \quad \text{during HP \& high-priority cluster} \tag{18}$$

$$N_{hl} g_{hl} \leq 1 - \frac{g_{hl} \eta_o W}{P_{Max}} \quad \text{during HP \& low-priority cluster} \tag{19}$$

Here N_{lh},N_{ll},N_{hh}, and N_{hl} are the number of mobiles in each of the different clusters. The parameters g_{lh}, g_{ll}, g_{hh}, and g_{hl} are defined as:

$$g_{ll} = \frac{\gamma_{ll}}{\gamma_{ll} + G_{ll}} \tag{20}$$

$$g_{lh} = \frac{\gamma_{lh}}{\gamma_{lh} + G_{lh}} \tag{21}$$

$$g_{hh} = \frac{\gamma_{hh}}{\gamma_{hh} + G_{hh}} \tag{22}$$

$$g_{hl} = \frac{\gamma_{hl}}{\gamma_{hl} + G_{hl}} \tag{23}$$

From the above inequalities (16-19), we can note the following:

1. If the SIR requirement, γ, is lower for a particular cluster then that cluster can support more users than a cluster of the same power level but with a higher target SIR. Following our specifications, since $\gamma_{ll} < \gamma_{lh}$, we get $N_{ll} > N_{lh}$. Similarly, we observe that $N_{hl} > N_{hh}$.
2. If the maximum transmit power requirement is higher for a particular cluster then it can support more simultaneous transmitters than can a cluster with the same SIR requirement but lower maximum transmit power. In other words, since $P_{Max} > P_{LPMax}$, $N_{hh} > N_{lh}$ and $N_{hl} > N_{ll}$.

As mentioned earlier, the modifications to the closed-loop power control for the LP clusters requires an increased target SIR during these slots. Therefore, we see that the SIR requirements can be ordered as either of the following:

$$\gamma_{hl} < \gamma_{hh} \leq \gamma_{ll} < \gamma_{lh} \tag{24}$$

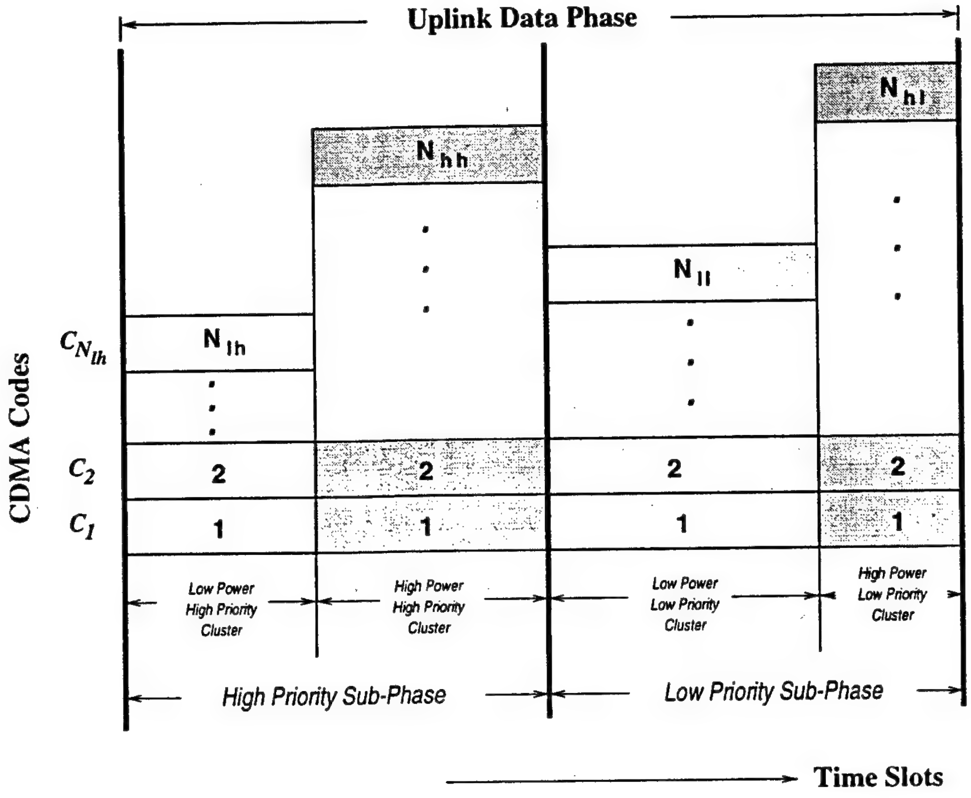
$$\gamma_{hl} < \gamma_{ll} \leq \gamma_{hh} < \gamma_{lh} \tag{25}$$

Combining this order and the two observations listed above, we get the following possible orders for the number of simultaneous transmitters that be supported during a slot in each cluster:

$$N_{lh} < N_{hh} \leq N_{ll} < N_{hl} \tag{26}$$

$$N_{lh} < N_{ll} \leq N_{hh} < N_{hl} \tag{27}$$

This result indicates that the scheduling algorithm must account for mobile battery power levels by not only prioritizing transmission but also adjusting the number of simultaneous transmission in each cluster so as to maintain QoS under power control modifications. The final ramifications of this scheduling process on the *uplink* phase is summarized in Figure 4.



Uplink Data Phase
CDMA Codes
$C_{N_{lh}}$
C_2
C_1
N lh
N hh
N ll
N hl
2
1
Low Power High Priority Cluster
High Power High Priority Cluster
Low Power Low Priority Cluster
High Power Low Priority Cluster
High Priority Sub-Phase
Low Priority Sub-Phase
Time Slots

Figure 4: General description of Uplink Phase with Battery Power Adaptation

By reducing the number of simultaneous transmitters in a CDMA-based system, we reduce interference which depends on simultaneous transmissions. In other words, if a particular channel - a slot/CDMA code combination - has a fewer number of users, it has lower interference, I_o. LP users can make use of this low interference and transmit at lower power levels to achieve the same error-rate as mobiles transmitting during HP clusters with higher power and more interference. In fact the scheduling algorithm can be adjusted to allow a lower error-rate for LP mobiles during their transmission cluster. Such an allocation reduces the number of retransmissions due to error and in turn conserves battery life.

There exist tradeoffs to scheduling a varying number of mobiles based on terminal battery life. There is a reduction in the capacity of the system. Since the number of simultaneous users in a particular slot could potentially be reduced, the total number of VC queues that a BS can support decreases with the number of LP mobiles. Also, it may appear that by reducing the number of simultaneous transmitters during LP slots will increase packet delay for LP users. In general terms, if the number of simultaneous users in a slot is reduced, LP users may have to wait additional slots *within the LP cluster* before transmitting. This increase in total packet delay is, however, offset by the high priority placed on LP users. Since LP clusters occur earlier in the frame, the LP mobiles have to wait a fewer number of slots *within the frame time* before transmission.

Additionally, to gain delay and throughput benefits for LP users, traffic from HP mobiles might suffer increased delay and lower throughput. Robust scheduling can accommodate HP mobiles so that they maintain a maximum packet delay and a minimum throughput based on the QoS requirements for the transmission stream. The consequence of the lower transmission priority is a possible degradation in the performance of the HP traffic when compared to LP mobiles. Further research is under progress to minimize the performance impact for HP mobiles.

5 Performance Analysis

Two sets of discrete-event simulations were performed to analyze the performance of the proposed algorithms. The difference between the two simulations was in the scheduler and the power control techniques. The first system employed a scheduler and power control mechanism that adapted to both terminal battery power and traffic priority. The second, on the other hand, only used traffic priority to schedule VC queues. In this section we generalize the assumptions used in our simulation runs and present their results.

5.1 Simulation Parameters

The parameters used in designing the discrete-event simulations are listed in Table 1. The simulation which adapted to terminal battery power analyzes a simple single-cell system in which there are N_1 mobiles. The analysis of the second system is performed using one BS but N_2 terminals. The reason for the two different number of mobiles is provided below.

The terminal mobility and channel propagation were modeled using the MADRAS (Mobility and Dynamic Resource Allocation Simulator) tool [19]. The free-flowing motion of the mobiles was generated with a truncated Gaussian speed distribution with a mean of μ_v, maximum of v_{max}, and minimum of v_{min}. The path loss models and shadow fading were implemented on the two-dimensional microcellular scope.

Mobiles were grouped as either LP or HP based on the ratio, r, of the number of LP mobiles to the total number of mobiles in the system. Each terminal had both a high-priority VC queue and a low-priority VC queue. We modeled the high-priority traffic with periodic packet arrivals, i.e. constant-bit-rate (CBR) traffic. For this traffic type, each terminal had to be assigned one slot in each uplink frame based on first-come-first-serve (FCFS) algorithm. The low-priority VC queues were formed at each terminal based on a Poisson arrival rate and scheduled for transmission according to shortest-job-first (SJF) algorithm. Two distinct schedulers were

Name	Value	Description
N_1	25	Number of mobiles in system with battery level adaptation
N_2	30	Number of mobiles in system without battery level adaptation
μ_v	25 km/hr	Mean speed of mobiles
v_{max}	33.33 km/hr	Maximum speed of mobiles
v_{min}	16.67 km/hr	Minimum speed of mobiles
r	0.3 & 0.5	Ratio of the number of LP mobiles to total mobiles
γ	-13 db	Target SIR for all mobiles
γ_{error}	-15 db	Minimum SIR for correct transmission
P_{LPMax}	1.5 mW	Transmit power-level for LP mobiles
P_{Max}	1.725 mW	Transmit power-level for HP mobiles
N_l	2	Number of simultaneous transmitters during LP clusters
N_h	3	Number of simultaneous transmitters during HP clusters
S	48	Number of slots in *Uplink* phase
U_l	2	Number of closed loop power control updates per LP slot
U_h	3	Number of closed loop power control updates per HP slot

Table 1: System Parameters

simulated: one which implemented our proposed algorithms and the other which prioritized transmission based only on traffic priority.

For simplification, the target SIR of γ was set to be the same for all VC queues and all clusters. Packets that were received below γ_{error} (based on a sample mean) were assumed corrupted and scheduled for retransmission. In the simulation that accounted for battery power levels, the transmit power for LP users was P_{LPMax} and P_{Max} for HP users. The HP transmit power level was adopted for all terminals in the second simulation - where no battery level priority was given to the N_2 mobiles. The scheduler that adapted to terminal battery assigned only N_l simultaneous transmitters during LP cluster slots and N_h HP users during their corresponding clusters. As shown earlier, $N_h > N_l$. The power control mechanism for this system updated transmit power levels U_l times during a LP slot and U_h times during a HP slot, where $U_h > U_l$. For the second scheduler, all slots could occupy N_h users and their transmit powers were similarly updated U_h times a slot.

In order to perform a fair comparison between the performance of the two systems, we selected N_1 and N_2 so that the two systems could support roughly the same capacity. The number of codes per slot available to LP users in the first system is N_l where as in the second system it is N_h. For high power users the number of codes per slot remains the same, N_h, in both systems. So the number of codes per slot available to HP and LP users is $N_l + N_h$ in the first system and $2N_h$ in the second, where $N_l + N_h < 2N_h$. Now we assume that roughly half the slots are used by HP users and the other half by LP users in both the systems. In order to make up for the disparity, we selected N_1 and N_2 so as to make the number of users per available code equal in the two systems, that is:

$$\frac{N_1}{N_l + N_h} = \frac{N_2}{2N_h} \tag{28}$$

For the case when $r = 0.5$, the assumption above concerning the number of slots allocated to LP and HP mobiles is fair for the system that does not prioritize based on battery power. This system will tend to allocate half the slots to LP mobiles and the other half to HP mobiles. In fact this is even a fair assumption for the adjusted system at low traffic load. However, in this system, at high packet arrival rates, the number of slots assigned to LP users increases. This implies a decrease in capacity; and hence the actual number of users that the system can support is less than the N_1 computed above for a fixed N_2. Thus the value of N_1 used in our simulation does not imply the two systems are operating under exactly the same capacity. It does, however, provide a closer measure of the actual capacity.

Since the capacity of the system changes with the traffic load, we will use the above N_1 as a rough means to achieve similar capacities in the two systems.

5.2 Performance Metrics

The performance of the two systems was studied by examining a set of parameters in varying traffic load for two different battery power assignments: 0.3 and 0.5. The traffic load was defined by the low-priority average packet generation rate at each mobile station. The parameters under analysis included the packet throughput, average packet delay, total power consumed per transmitted packet, and finally the packet error rate. Each of these four metrics were examined first for all LP mobiles in the system and then just for the HP mobiles. For the simulation that used our proposed algorithms, the throughput for LP mobiles, Γ_l, and throughput for HP mobiles, Γ_h, were computed as:

$$\Gamma_l = \frac{X_{ls}}{SN_l} \tag{29}$$

$$\Gamma_h = \frac{X_{hs}}{SN_h} \tag{30}$$

where X_{ls} and X_{hs} represent the total number of packets transmitted successfully per frame by LP and HP mobiles respectively. In a similar fashion, the throughput calculations for the mobiles in the system without battery level adaptation were:

$$\Gamma_l = \frac{X_{ls}}{SN_h} \tag{31}$$

$$\Gamma_h = \frac{X_{hs}}{SN_h} \tag{32}$$

Note that N_h is used in both equations (31) and (32). This is because the number of simultaneous users per slot was defined to the same for all mobiles in this system.

The average packet delay was the number of slots a packet had to wait before transmission. The ratio of the total power consumed by all terminals during the simulation (both in transmit and standby mode) to the total number of packets transmitted yielded a measurement for the power consumption of the system. Finally as a means to show the transmission reliability of the two systems, we determined the error rate during the simulation as the total number of packets transmitted unsuccessfully to the total number of packets transmitted.

5.3 Simulation Results

The packet throughput, delay, power efficiency, and error rate for both the LP and HP mobiles were computed during both the simulations. The results are presented in Figures 5-6 for two different ratios, $r = 0.3$ and $r = 0.5$. The measurements are shown separately for LP and HP mobiles for increasing packet arrival rate.

Figure 5(a) shows the improvement of packet throughput at LP mobiles when operating in a battery adaptive system. First consider the system without battery adaptation. In this system, we observe by comparing Figures 5(a) and (b) that the throughput for both LP and HP users is almost identical when the number of HP and LP mobiles is the same. This is not surprising since the system makes no scheduling decisions based on battery power levels. When there are more HP users, i.e. when $r = 0.3$, the throughput for the HP users is higher at each packet arrival rate since more HP packets are transmitted. As the traffic load increases, the LP and HP throughput – for both ratios – increase due to increased packet arrivals and transmissions. At heavy traffic load we see that the throughput for the HP users in this system decreases when $r = 0.3$ whereas it continues to increase for $r = 0.5$. The large number of HP mobiles eventually results in longer buffered queues during heavy traffic situations at HP mobiles. Since there are fewer LP mobiles in this scenario, the total number of packets buffered at their queues is less. This disparity causes the dip in the throughput performance for HP mobiles without battery adaptation when $r = 0.3$.

The throughput results for the scheduler which incorporates battery power show that the LP users' throughput is higher despite increasing traffic. This occurs because of the high-priority given to packets from LP terminals. When comparing the performance of this system for different ratios of LP mobiles, we see that as the number of LP mobiles increases, i.e. $r = 0.5$, the throughput of the system is higher. The reason for this lies in the increased number of packets transmitted by all the LP mobiles when r is larger. A larger number of transmitted packets in turn produces a higher throughput.

The high-priority of the LP terminals is partly responsible for the reduced throughput performance of HP

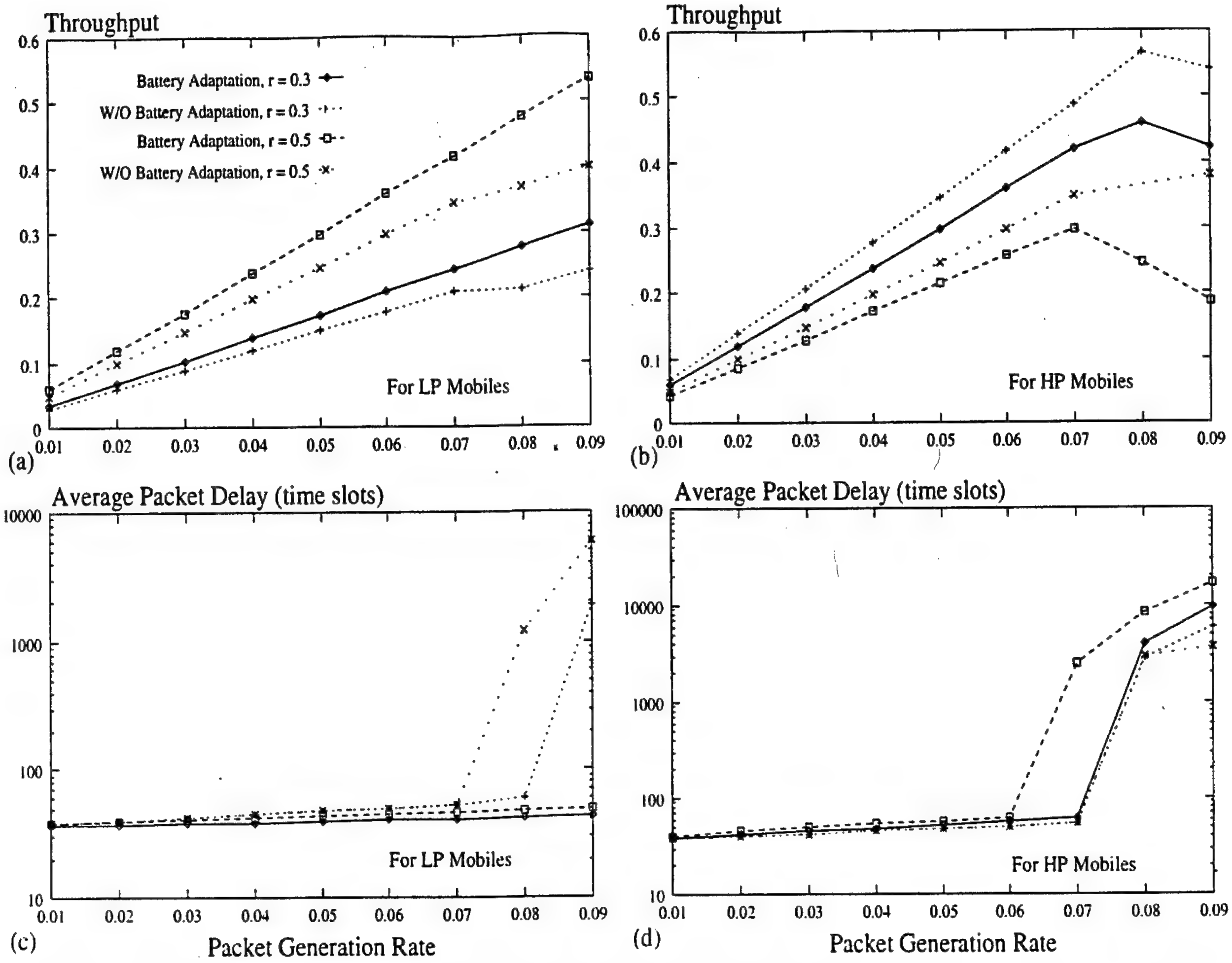


Figure 5: Throughput for (a) low-power mobiles (b) high-power mobiles and Packet Delay for (c) low-power mobiles (d) high-power mobiles. r is the ratio of low-power mobiles to the total number of mobiles in the system

users under heavy load. We see from Figure 5(b) that the throughput of HP users is lower with our proposed battery adaptation for both ratios r. In fact as opposed to the curves for the adjusted system in Figure 5(a), the throughput for HP mobiles decreases during heavy traffic. The lower HP throughput results in part because of the simplified scheduling of low-priority traffic which allots only leftover slots of low-priority & HP mobiles. At high traffic the number of leftover slots reduces as these slots are assigned to LP mobiles thus reducing the throughput for the HP terminals. When comparing the throughput for varying ratios, we see that the throughput performance for HP mobiles is lower when there are greater number LP mobiles. This again results because more LP mobiles cause less leftover slots which in turn reduces the number of packets transmitted from HP terminals.

The results for packet delay in Figures 5(c) and (d) also indicate an improvement for LP mobiles operating under battery level adaptation. For the system that does not prioritize between LP and HP mobiles, we see when comparing the appropriate curves in Figures 5(c) and (d) that the delay performance of the two battery classes are relatively similar. The delay at both LP and HP mobiles starts to grow larger as the traffic increases. This occurs because as the arrival rate of the packets increases, the buffered queue gets larger and subsequently so does the time that a buffered packet has to wait before transmission. In the system with battery level adaptation, the packet delay as shown in Figure 5(c) for LP users remains low even for high packet arrival rates when compared to the system without battery level adjustments. The difference between the average packet delay of the two systems gets larger for LP mobiles with increasing traffic load. At high traffic load, the battery adaptive system starts to allocate more and more channels to LP mobiles and thereby counters the effect of the large packet arrival rate at its VC queues. This way the scheduler maintains a relatively small number of buffered packets which then reduces the delay for the packets that are buffered. These lower delay measurements for the LP users are seen for both values of r. For the case when r is smaller, i.e. when there are fewer number of LP mobiles, we see that the delay is smaller for the system with battery adaptation. A fewer number of LP users means that the scheduler can assign more packets during a fixed number of slots. This way the buffered queue is smaller when there are fewer number of LP mobiles and thus the delay on the buffered packets is reduced.

The delay measurements for the HP users in Figure 5(d), on the other hand, indicate a different result. For all traffic loads, we see that in the system with the battery adaptation there is a greater delay on VC queues originating from HP terminals. Note that at lower traffic rates the average packet delay of the adjusted system is closer to that of the system without battery adaptation. This difference gets significantly larger at very high traffic rates since at these loads LP users are assigned most of the available low-priority slots. The direct result of this preference for LP transmissions is that packets at HP terminals must wait longer to be assigned a slot at high traffic load. This wait increases as the number of LP mobiles increases, that is when $r = 0.5$. Thus the transmission priority given to LP mobiles starts to deteriorate the performance of HP mobiles when the packet arrival rate increases.

One way to improve the performance of the HP mobiles in the heavy traffic situations is to develop a scheduling algorithm during HP clusters specifically designed to maintain a minimum throughput and maximum packet delay. One possibility is using a scheduler that prioritizes high-power users once in every few frames. Other such techniques for improving HP performance during heavy traffic are currently being studied and will be the focus of future work.

As shown in Figures 6(a) and (b), battery power adaptation provides energy efficient transmissions for LP terminals. First, we observe in Figure 6(a) that the total power consumed per transmitted packet for LP users under battery power adjustments is less than that for the other system at both ratios. This occurs because of the energy conserved at LP mobiles due to lower transmit power levels and less frequent power control updates which also exhaust battery supply. The power expended in the system without battery adaptation is similar for both LP and HP mobiles. The reason for this again lies in the lack of differences between LP

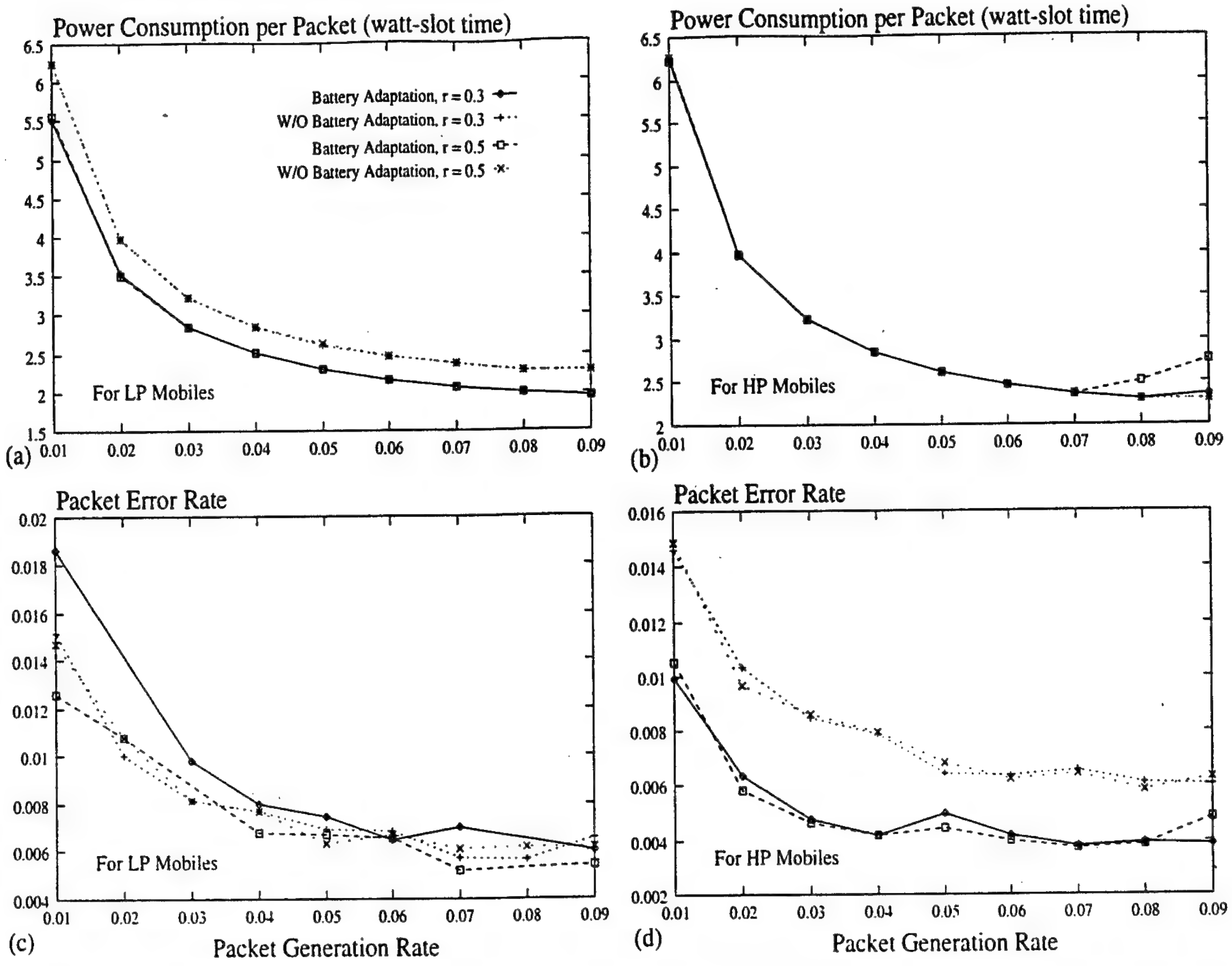


Figure 6: Power Consumption per Packet for (a) low-power mobiles and for (b) high-power mobiles, and Packet Error Rate for (c) low-power mobiles and for (d) high-power mobiles. r is the ratio of low-power mobiles to the total number of mobiles in the system.

and HP transmissions in this system; LP and HP mobiles transmit at the same power level and are monitored with equal number of power control updates. The power consumed per transmitted packet is higher during light traffic loads because terminals spend a significant portion of the time in the idle mode since they have a relative small number of packets to transmit. Due to the larger amount of power expended at the stand-by level and the few number of transmitted packets, the power efficiency at low traffic load is measured to be larger than during heavy traffic.

In Fig. 6(b), we see that HP users expend almost the same power per packet transmission in both systems for low to moderate traffic load. HP mobiles in the system with battery adaptation transmit fewer number of packets as indicated in Figure 6(a) and therefore consume less power. Since the transmit power level is the same for the two systems, the power consumed per transmission is thus the nearly the same in both the systems under analysis. This, however, is not the case at heavier traffic loads when HP users with battery adaptation – for both ratios – expend more power per packet than do HP users in the other system. In heavy traffic, HP terminals in the battery adaptive system spend a great deal of time in the idle mode; they transmit very few packets and yet expend power transmitting at the standby level. This then yields a large amount of power expended for each transmitted packet and thus causes a degradation in the power efficiency at heavy traffic load.

The results of the adjustments on the packet error rate can be seen in Figures 6(c) and (d). These figures provide a means to show that our algorithms did not degrade the desired error rate as compared to the system without the power control adjustments. As we see in Figure 6(c), the error rate, measured as the ratio of the number of unsuccessful packet transmissions to the total number of packets transmitted is similar for LP users in the two systems. We also observe that the error-rate performance for the system without battery adaptation is almost the same for both ratios and both battery power levels. This is because the both LP and HP users in this system were given the same transmit power levels, target SIR, and number of simultaneous transmitters. The power control adjustments that we implemented, i.e. reducing the transmit power level and the number of simultaneous transmitters, result in a system that produced very similar packet error rates for LP mobiles, as shown by the curves in Figure 6(c). The packet error rate for a larger number of LP mobiles is lower since in this scenario a larger number of packets are transmitted and a greater number of slots are assigned to LP terminals. The packet error rates for HP mobiles in the battery adaptive system shown as the bottom set of curves in Figure 6(d) are lower for both ratios. Since HP mobiles transmit less often in the adjusted system, the total number of transmitted packets reduces and so does the packet error rate.

Using our proposed algorithms, we see that terminals low on battery power can transmit more packets per frame and thereby achieve greater throughput. Additionally, the packet delay for LP users is reduced when operating under the battery adaptive system. The improvement in throughput and packet delay is beneficial since it allows greater communications capability before complete battery loss. The power efficiency gained by the power control adjustments adds to this capability by extending battery usage over time. Finally, as shown by the packet error rate measurements, these benefits are gained without sacrificing the QoS criterion on the target error rate.

6 Summary

The paper focused on techniques to help improve the transmission capabilities of low-power mobiles in a hybrid CDMA/TDMA system. The time-division properties of the system were exploited to schedule transmission times to users based not only on their traffic priority but also terminal battery status. The clustering phenomena that resulted from the proposed scheduling modifications was then used as the basis to modify CDMA power control algorithms. The power control adjustments were geared to adapt to the battery status of each cluster and hence of the serving mobiles. A discrete-event simulation was performed and its results presented. The analysis demonstrated a performance enhancement for low-power terminals operating under our proposed algorithms. The increased throughput, reduced average packet delay, and power efficiency were shown to be gained without sacrificing the error rate QoS criterion for the low-power terminal.

References

[1] K. Pahlavan and A. H. Levesque, *Wireless Information Networks.* A Wiley-Interscience Publication, 1995.

[2] M. Naghshineh (Guest Ed.), "Special issue on Wireless ATM," *IEEE Personal Communications*, vol. 3, Aug. 1996.

[3] S. Singh, "Quality of Service guarantees in mobile computing," *Computer Communications*, vol. 19, pp. 359–371, Apr. 1996.

[4] K. M. Sivalingam, M. B. Srivastava, P. Agrawal, and J.-C. Chen, "Low-power Access Protocols Based on Scheduling for Wireless and Mobile ATM Networks," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, Oct. 1997. (To Appear).

[5] K. M. Sivalingam, M. B. Srivastava, and P. Agrawal, "Low power link and access protocols for wireless multimedia networks," in *Proc. IEEE Vehicular Technology Conference*, (Phoenix, AZ), pp. 1331–1335, May 1997.

[6] P. Agrawal, J.-C. Chen, and K. M. Sivalingam, "Battery power consumption based analysis of MAC protocols for wireless multimedia networks," tech. rep., Washington State University, Pullman, WA, July 1997.

[7] M. Stemm, P. Gauthier, D. Harada, and R. H. Katz, "Reducing power consumption of network interfaces in hand-held devices," in *Proc. of 3rd International Workshop on Mobile Multimedia Communications (MoMuc-3)*, (Princeton, NJ, USA), Sept. 1996.

[8] M. Zorzi and R. R. Rao, "Error control and energy consumption in communications for nomadic computing," *IEEE Transactions on Computers*, vol. 46, pp. 279–289, Mar. 1997.

[9] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication.* Addison-Wesley Publishing Company, 1995.

[10] A. Sampath, S. Kumar, and J. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. of IEEE PIMRC*, Sept. 1995.

[11] R. Prasad, *CDMA for Wireless Personal Communications.* Artech House, 1996.

[12] T. S. Rappaport, *Wireless Communications: Principles and Practice.* Prentice Hall, 1996.

[13] R. C. Dixon, *Spread Spectrum Systems, 2nd ed.* Wiley, 1984.

[14] D. Goodman, *Wireless Personal Communications Systems.* Addison-Wesley Publishing Company, 1997.

[15] TIA/EIA, "Mobile station – base station compatibility standard for dual-mode wideband spread spectrum cellular system." TIA/EIA Interim Standard-95, July 1993.

[16] M. A. Arad and A. Leon-Garcia, "Scheduled CDMA: a hybrid multiple access for wireless ATM networks," in *Proc. of IEEE PIMRC*, pp. 913–917, 1996.

[17] M. D. Prycker, *Asynchronous Transfer Mode: Solution for Broadband ISDN.* Prentice Hall, 3 ed., 1995.

[18] A. M. Viterbi and A. J. Viterbi, "Erlang capacity of a power controlled CDMA system for portable applications," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 6, pp. 892–900, 1993.

[19] K. G. Chen, *Integrated dynamic radio resource management of wireless communication systems.* MS Thesis, Rutgers University, May 1996.

To APPEAR IN IEEE ICUPC. 98, OCT. 1998

A Battery Power Level Aware MAC Protocol for CDMA Wireless Networks

Shalinee Kishore[1], Jyh-Cheng Chen[2], Krishna M. Sivalingam[3], and Prathima Agrawal[4] *

[1]WINLAB, Rutgers University, Piscataway, NJ 08855

[2]Department of Electrical & Computer Engineering, State University of New York at Buffalo, Buffalo, NY 14260

[3]School of Electrical Engineering & Computer Science, Washington State University, Pullman, WA 99164

[4]Networked Computing Technology Department, AT&T Labs, Whippany, NJ 07981

Abstract - *A general constraint common to many wireless networks lies in the short lifetime of mobile terminal batteries. Energy efficient protocols that adapt to terminal battery power level can be used to reduce the effects of this limitation. This paper addresses such power-adaptive algorithms in medium access control (MAC) protocols and in the power control mechanisms for hybrid CDMA/TDMA wireless networks. Our access protocol dynamically schedules CDMA channels to mobiles based on their traffic requests and battery power levels. This technique assigns mobiles with similar traffic requests and battery power levels to one or more slots. Discrete-event simulation has been used to demonstrate that the proposed techniques indeed provide low-power mobiles with increased throughput and reduced latency while reducing power usage.*

1 Introduction

A general constraint on wireless communications lies in the short lifetime of mobile terminal batteries. Due to this limitation, it has been proposed that low-power design should also be a crucial consideration in designing all layers of the protocol stack for wireless networks [1]. Typically, power conservation is considered at the hardware layers within the mobile terminal or in error control [2, 3]. In addition to hardware considerations, the wireless infrastructure should use information about each user's battery level and adapt network operation accordingly. As discussed in [1], the CPU, the transmitter, and the receiver are the major consumers of battery power at the mobile terminal for access protocol activities.

The performance of standard, multi-rate CDMA systems, such as the one described in [4], is restricted by the power-budgets of low-power (LP) mobiles, i.e. those mobiles whose battery supply is the lowest. Additionally such systems require LP mobiles to transmit data at the maximum transmit power available, thereby straining their battery supply. To reduce the burden placed on LP users, we propose in this paper the use of a scheduled CDMA system. In this scheme, transmissions from CDMA-based mobiles are coordinated from a central scheduler located at the serving basestation (BS). Such systems have been developed in previous work [5]- [6] but not with aims to improve power efficiency.

The benefits of adopting a scheduled system, such as the Energy-Conserving Medium Access Protocol (EC-MAC) [7], have been studied in [1]. EC-MAC relies – for energy conservation reasons – on scheduling algorithms to assign transmission times to mobiles.

The scheduling scheme proposed aims to reduce the simultaneous interference based on the knowledge of the battery power level of each user. The scheduling algorithm is explicitly suited to benefit LP users in the hybrid system. The algorithm prioritizes LP mobile traffic to be scheduled for transmission separate from high-power (HP) mobile traffic. At the same time, the scheduler also considers the registered priority of the traffic – in terms of the desired transmission rates, error rates and delay sensitivity.

One ramification of this scheduling algorithm is that LP users are assigned adjacent transmission slots and hence are "clustered" in time. The modifications proposed here are a result of this "clustering" phenomena and affect the power control algorithm. All systems, particularly CDMA-based ones, use power control algorithms which – among other things – govern the transmission power levels of the mobiles. The scheduler must then dynamically adjust the number of simultaneous transmitters, i.e. number of users assigned to a particular slot, so that LP "clusters" contain slots with lower interference. Due to the reduced interference, lower transmit powers can be assigned to LP users without effecting the QoS (Quality of Service) in terms of the target bit error rate. of the transmitted data stream.

2 System Description

The paper considers a single-cell environment. Each mobile in the cell continuously communicates with the BS using a random sequence. The chip rate for all users is fixed and the entire system bandwidth, W, is used by all users. For the hybrid system under consideration here, time is divided into equal-length *slots*. Transmission slots are assigned to each user by a centralized scheduler at the BS. During its assigned slot(s), the mobile must power up its transmitter to a specified transmit power level and send out its digital stream which maybe buffered in a queue. When not transmitting information, i.e. when a mobile is not assigned to the current slot, the terminal uses its code to communicate with the BS at a lower power level for synchronization purposes. In the analysis here, we will assume that this synchronization contributes negligible interference to other transmitting mobiles.

The hybrid system discussed here relies on a scheduling mechanism to coordinate user access to the transmission slots. The MAC protocol is derived from the EC-MAC protocol defined in [7]. Data sent from or to the mobile at this layer is in units of *packets* which we assume occupies exactly one slot. Each packet has a particular transmission rate and priority assigned to it. The frame structure of the EC-MAC protocol used here is shown in Figure 1. Transmission in the frame is divided into frames which is further divided into phases. Of primary concern to us is first the *request/update and new user* phase in which registered mobiles transmit their current queue status, battery power level, etc. to the BS. Also during this phase, new users register with the BS. Next is the *downlink broadcast* phase when the BS broadcasts data, acknowledgments, scheduling information, and transmit power level assignments that all mobiles need to received. Then finally comes the *downlink unicast/multicast and uplink* phase. During this phase the BS unicasts/multicasts data to different sets of users. At the same time, registered mobiles transmit their buffered packets to the BS during scheduled

*The authors can be reached at kishore@ece.rutgers.edu, jcchen@eng.buffalo.edu, krishna@eecs.wsu.edu, and pa@research.att.com. Part of the research was supported by Air Force of Scientific Research grant F-49620-97-1-0471, and by AT&T Labs, Whippany, NJ.

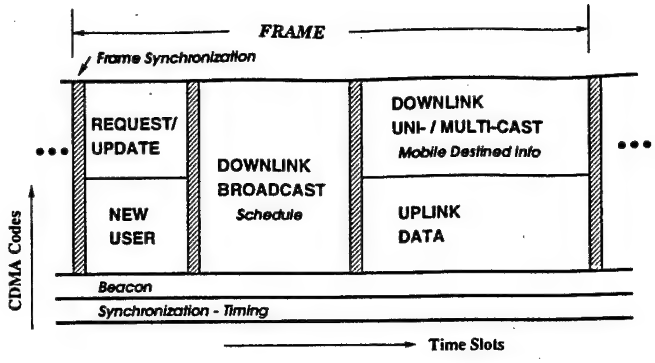


Fig. 1. Frame structure of the multiple access protocol showing the various sub-phases.

slots.

The power control mechanism presented here for a hybrid CDMA/TDMA systems is responsible for assigning transmit power levels to each mobile in the cell. These power levels are broadcast to each mobile during the *downlink broadcast* phase, as indicated earlier. The mobiles in turn then adjust their powers to the specified levels during their assigned slot(s) of the *uplink* phase. These transmit power values are determined at the BS based on the minimum rate and SIR requirements of all simultaneously transmitting users. Thus, during a slot, users transmitting at these given powers can achieve (at the least) a specific data rate and minimum SIR performance. This minimum SIR maps to a specific error rate which in turn can be used to maintain QoS guarantees for all users [8]. The optimal power allocation for each user, P_i, in the current slot is found as:

$$P_i^* = \eta_o W \left(\frac{\gamma_i R_i}{W + \gamma_i R_i}\right)\left[1 - \sum_{j=1}^{N} \frac{\gamma_j R_j}{W + \gamma_j R_j}\right]^{-1} ,\forall \text{ i} = 1..\text{N} \qquad (1)$$

where η_o is the one-sided power spectral density of additive white Gaussian noise, γ_i and R_i are the minimum SIR and rate requirements, and N is the number of users assigned to the current slot. The existence of this optimal solution depends on the constraints on P_i. If the maximum power constraint on user i is P_{Max_i} (i.e., $0 \leq P_i \leq P_{Max_i}$), then it is required that:

$$\sum_{j=1}^{N} \frac{\gamma_j R_j}{W + \gamma_j R_j} \leq 1 - \frac{\eta_o W}{\min_i \left[P_{Max_i}\left(\frac{W}{\gamma_i R_i} + 1\right)\right]} \qquad (2)$$

The details of the derivations of these power assignments can be found in [4].

3 Proposed Algorithms

The techniques proposed to adapt network operations to lower-power terminals are described in this section. The algorithm operates under the assumption that the BS has knowledge of the battery power level of all mobiles in its coverage area. Thus the first requirement of our algorithm is a periodic battery power level update when mobiles transmit their current battery supply to their serving BS. Based on a simple threshold comparison, the BS groups mobiles of similar power levels together. Our description here and the system simulation assumes that mobiles are classified into two types: HP and LP (high-power and low-power). The algorithm can easily be generalized to more number of discrete power levels. The packet queues at each mobile are also classified based on their QoS priority. Here we assume two priority levels – high-priority and low-priority for all users in the system. More specifically, high-priority traffic has the same rate and SIR requirements (R_h and γ_h, respectively) and is delay-sensitive. All low-priority traffic also has the same rate and SIR requirements (R_l and γ_l) but is delay tolerant. Again, the algorithm can be easily generalized to more priority levels.

3.1 Scheduling Adjustments

The scheduling algorithm allocates slots in the *uplink data phase* of the MAC frame to each packet queue based on terminal battery power level and priority. One of the aims of our algorithm is to reduce the latency at the LP mobile packet queues. To do this, we must schedule transmission of packets from these LP mobiles as early as possible in the *uplink* phase.

An added dimension to this scheduling lies in the prescribed priority of each packet queue. The higher priority classifications maintain their own minimum delay, SIR, and error rate requirements. These specifications are used to quantify the QoS of that connection.

These two considerations of battery power levels and packet queue priority results in the following general scheduling scheme. The BS breaks down the *uplink* phase into four intervals or "clusters." The "clusters" are defined by the combination of the two scheduling parameters: LP & high-priority, HP & high-priority, LP & low-priority, and finally HP & low-priority. If traffic priority was the only scheduling consideration, then BSs would allocate the first available slots to high-priority queues and then the remaining slots would be assigned to low-priority, delay-tolerant queues. With the knowledge of mobile battery status, we propose that the high-priority slots should first be assigned to LP users thus forming a LP & high-priority cluster at the start of the *uplink* phase. The HP & high-priority mobiles will then be given the next cluster, followed by LP & low-priority and then the HP & low-priority.

3.2 Power Control Adjustments

As a consequence of our scheduling adjustments, we observe that low power users transmit during two specific time intervals of the *uplink* phase. Since the BS handles the scheduling of the uplink slots, it has exact knowledge about the start and end times of these intervals. Due to this knowledge, we now require the BS dynamically adjusts its power specifications to adapt to the presence of *all* LP users during a particular cluster or contiguous slots, i.e. the BS reduce its power control requirements during that segment of time.

Let P_{LPMax} represent the maximum power level that a LP mobile can transmit at during its cluster. Note that if P_{Max} is the maximum power at which all mobiles transmitted before battery power adaptation, then we require that $P_{LPMax} < P_{Max}$. Additionally, note that the HP users can still transmit at P_{Max} during HP clusters since they are not under the same power budgets.

Our adjusted power allocation is similar to equation (1) with the additional power constraints:

$$0 < P_i \leq P_{LPMax} \ \forall \text{ mobile } i \in \text{LP Clusters} \qquad (3)$$

$$0 < P_j \leq P_{Max} \ \forall \text{ mobile } j \in \text{HP Clusters} \qquad (4)$$

Now, note that due to the traffic priority classification, users in a particular slot share the same minimum rate requirement and minimum SIR measurement. As indicated above R_h is the minimum rate requirement during high-priority slots and R_l is the minimum rate requirement during low-priority slots. In the same manner, γ_h and γ_l are the minimum SIR requirements for the two traffic types. Due to its priority, we assume here $\gamma_h R_h > \gamma_l R_l$. Using the solution (1) and the above transmit power limitation, the constraint in (2) can then be simplified for each cluster.

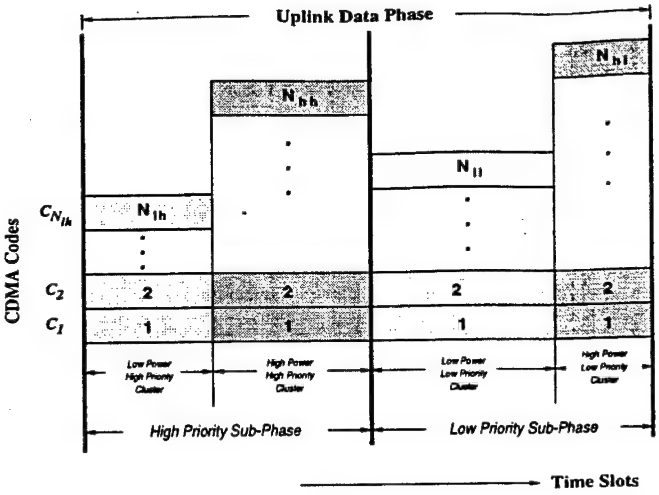

Uplink Data Phase
CDMA Codes
$C_{N_{lh}}$
C_2
C_1
N lh
N hh
N ll
N hl
2
1
Low Power High Priority Cluster
High Power High Priority Cluster
Low Power Low Priority Cluster
High Power Low Priority Cluster
High Priority Sub-Phase
Low Priority Sub-Phase
Time Slots

Fig. 2. General description of Uplink Phase with Battery Power Adaptation

Let N_{Lh}, N_{Hh}, N_{Ll}, and N_{Hl} be the number of mobiles assigned to each slot during the LP & high-priority, HP & high-priority, LP & low-priority, and HP & low-priority clusters, respectively. Since $\gamma_h > \gamma_l$, $R_h > R_l$, and $P_{Max} > P_{LPMax}$, it can easily be shown using equation (2) that the number of simultaneous transmitters that may be supported during a slot in each cluster are related as either of the following:

$$N_{Lh} < N_{Hh} \leq N_{Ll} < N_{Hl} \tag{5}$$

OR

$$N_{Lh} < N_{Ll} \leq N_{Hh} < N_{Hl} \tag{6}$$

This result indicates that the scheduling algorithm must account for mobile battery power levels by not only prioritizing transmission but also adjusting the number of simultaneous transmission in each cluster so as to maintain QoS under power control modifications. The final ramification of this scheduling process on the *uplink* phase is summarized in Figure 2.

By reducing the number of simultaneous transmitters in a CDMA-based system, we reduce the total interference for each transmitting user. LP users with a particular traffic requirement can make use of this low interference and transmit at lower power levels to achieve the same error-rate.

There exist tradeoffs to scheduling a varying number of mobiles based on terminal battery life. There is a reduction in the capacity of the system. Since the number of simultaneous users in a particular slot could potentially be reduced, the total number of information streams that a BS can support decreases with the number of LP mobiles. Also, it may appear that by reducing the number of simultaneous transmitters during LP slots will increase packet delay for LP users. This increase in total packet delay is, however, offset by the high priority placed on LP users.

Additionally, to gain delay and throughput benefits for LP users, traffic from HP mobiles might suffer increased delay and lower throughput. Robust scheduling can accommodate HP mobiles so that they maintain a maximum packet delay and a minimum throughput based on the QoS requirements for the transmission stream. Due to the elimination of LP users during HP transmissions, it maybe possible for HP users to improve their rate performance, i.e. exceed the minimum rate requirements. This can help offset the lower throughput that results at the higher layer due to LP packet priority. Further research is under progress to quantify and minimize the performance of HP mobiles.

Name	Value	Description
N_1	25	No. of mobiles in system with our scheme
N_2	30	No. of mobiles in system w/o our scheme
μ_v	25 km/hr	Mean speed of mobiles
v_{max}	33.33 km/hr	Maximum speed of mobiles
v_{min}	16.67 km/hr	Minimum speed of mobiles
r	0.3 & 0.5	Ratio of No. of LP mobiles to total mobiles
γ	-13 db	Target SIR for all mobiles
γ_{error}	-15 db	Minimum SIR for correct transmission
P_{LPMax}	1.5 mW	Transmit power-level for LP mobiles
P_{Max}	1.725 mW	Transmit power-level for HP mobiles
N_l	2	No. of simult. X-mitters during LP clusters
N_h	3	No. of simult. X-mitters during HP clusters
S	48	No. of slots in *Uplink* phase

Table 1. System Parameters

4 Performance Analysis

Two sets of discrete-event simulations were performed to analyze the performance of the proposed algorithms. The difference between the two simulations was in the scheduler and the power control techniques. The first system employed a scheduler and power control mechanism that adapted to both terminal battery power and traffic priority. The second, on the other hand, only used traffic priority to schedule packet transmissions. In this section we generalize the assumptions used in our simulation runs and present their results.

4.1 Simulation Parameters

The parameters used in designing the discrete-event simulations are listed in Table 1. The simulation which adapted to terminal battery power analyzes a simple single-cell system in which there are N_1 mobiles. The analysis of the second system is performed using one BS but N_2 terminals. The reason for the two different number of mobiles is discussed later.

The terminal mobility and channel propagation were modeled using the MADRAS (Mobility and Dynamic Resource Allocation Simulator) tool [9]. The free-flowing motion of the mobiles was generated and path loss models and shadow fading were implemented on the two-dimensional micro-cellular scope.

Each terminal had both a high-priority and low-priority packets. Thus each mobile maintains two packet queues for transmissions. We modeled the high-priority traffic with periodic packet arrivals, i.e. constant-bit-rate (CBR) traffic which had to be assigned one slot in each uplink frame based on first-come-first-serve (FCFS) algorithm. The low-priority packet queues were formed at each terminal based on a Poisson arrival rate and scheduled for transmission according to shortest-job-first (SJF) algorithm.

The HP transmit power level was adopted for all terminals in the second simulation - where no battery level priority was given to the N_2 mobiles. Note also that in the simulation, we accounted for the power transmitted to maintain synchronization during unassigned slots, i.e. the stand-by power. The scheduler that adapted to terminal battery assigned only N_l simultaneous transmitters during LP cluster slots and N_h HP users during their corresponding clusters. For the second scheduler, all slots could occupy N_h users.

In order to perform a fair comparison between the performance of the two systems, we selected N_1 and N_2 so that the two systems could support roughly the same capacity. For high power users the number of channels per slot remains the same, N_h, in both systems. So the number of channels per slot available to HP and LP users is $N_l + N_h$ in the first system and $2N_h$ in the second, where $N_l + N_h < 2N_h$. Now we assume that roughly half the slots are used by HP users and the other half by LP users in both the systems. In order to make up for the disparity, we selected N_1 and N_2 so as to make the number of users per available code equal in the two systems, that is: $\frac{N_1}{N_l+N_h} = \frac{N_2}{2N_h}$.

For the case when $r = 0.5$, the assumption above concerning the number of slots allocated to LP and HP mobiles is fair for the system that does not prioritize based on battery power. This system will tend to allocate half the slots to LP mobiles and the other half to HP mobiles. In fact this is even a fair assumption for the adjusted system at low traffic load. However, in this system, at high packet arrival rates, the number of slots assigned to LP users increases. This implies a decrease in capacity; and hence the actual number of users that the system can support is less than the N_1 computed above for a fixed N_2. Thus the value of N_1 used in our simulation does not imply the two systems are operating under exactly the same capacity. It does, however, provide a closer measure of the actual capacity.

4.2 Performance Metrics

The performance of the two systems was studied by examining a set of parameters in varying traffic load for two different battery power assignments: $r = 0.3$ and $r = 0.5$ for both LP and HP users. The traffic load was defined by the low-priority average packet generation rate at each mobile station. For the simulation that used our proposed algorithms, the throughput for LP mobiles, Γ_l, and throughput for HP mobiles, Γ_h, were computed as: $\Gamma_l = \frac{X_{ls}}{SN_l}$ and $\Gamma_h = \frac{X_{hs}}{SN_h}$. Here X_{ls} and X_{hs} represent the total number of packets transmitted successfully per frame by LP and HP mobiles respectively. In a similar fashion, the throughput calculations for the mobiles in the system without battery level adaptation were $\Gamma_l = \frac{X_{ls}}{SN_h}$ and $\Gamma_h = \frac{X_{hs}}{SN_h}$.

The average packet delay was the number of slots a packet had to wait before transmission. The ratio of the total power consumed by all terminals during the simulation (both in transmit and standby mode) to the total number of packets transmitted yielded a measurement for the power efficiency. Finally we determined the error rate during the simulation as the total number of packets transmitted unsuccessfully to the total number of packets transmitted.

4.3 Simulation Results

The results for the packet throughput, delay, power efficiency, and error rate for both the LP and HP mobiles are presented in Figures 3-4 for two different ratios, $r = 0.3$ and $r = 0.5$. The measurements are shown separately for LP and HP mobiles for increasing packet arrival rate.

First consider the throughput for the system without battery adaptation. In this system, we observe by comparing Figures 3(a) and (b) that the throughput for both LP and HP users is almost identical when the number of HP and LP mobiles is the same. This is not surprising since this system makes no scheduling decisions based on battery power levels. When there are more HP users, i.e. when $r = 0.3$, the throughput for the HP users is higher at each packet arrival rate since there are more HP users and hence more HP packets are transmitted. As the traffic load increases, the LP and HP throughput – for both ratios – increase since more packets are arriving at their queues and hence are being transmitted. At heavy traffic load we see that the throughput for the HP users in this system decreases when $r = 0.3$ whereas it continues to increase for $r = 0.5$. The large number of HP mobiles eventually results in longer buffered queues during heavy traffic situations at HP mobiles. Since there are fewer LP mobiles in this scenario, the total number of packets buffered at their queues is less. This disparity causes the dip in the throughput performance for HP mobiles without battery adaptation when $r = 0.3$.

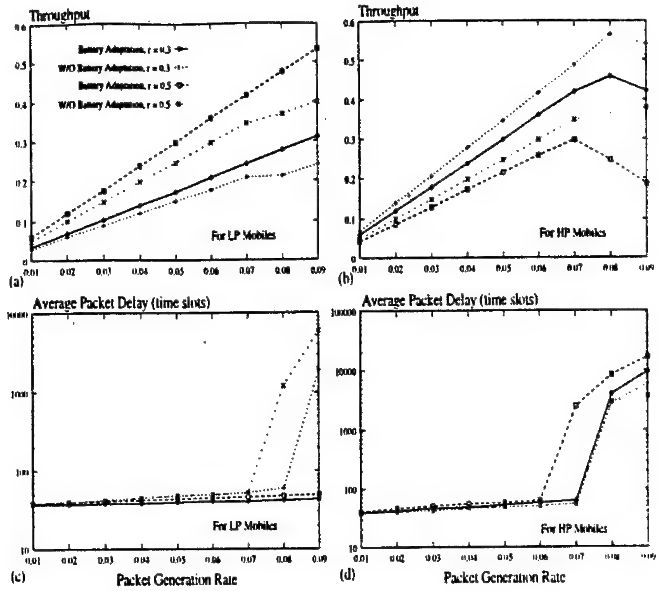


Fig. 3. Throughput for (a) low-power mobiles (b) high-power mobiles and Packet Delay for (c) low-power mobiles (d) high-power mobiles. r is the ratio of low-power mobiles to the total number of mobiles in the system

The throughput results for the scheduler that incorporates battery power show that the LP users' throughput is higher and remains higher with increasing traffic when compared to the previous system. Since LP terminals are given a preference when scheduling, buffered packets at these terminals are transmitted more frequently than the system that does not give LP users this priority. Since there are more LP mobiles when $r = 0.5$, the number of packets transmitted is higher and thus the throughput of the system is higher.

In the battery-adaptive system, the high-priority of the LP terminals is partly responsible for the throughput performance of HP users. We see from Figure 3(b) that the throughput of HP users is lower with our proposed battery adaptation for both ratios r. In fact as opposed to the curves for the adjusted system in Figure 3(a), the throughput for HP mobiles decreases during heavy traffic. The lower HP throughput results in part because of the simplified scheduling of low-priority traffic which allots only leftover slots of low-priority & HP mobiles. At high traffic or when there are more LP users ($r = 0.5$), the number of leftover slots reduces as these slots are assigned to LP mobiles thus reducing the throughput for the HP terminals.

The results for packet delay in Figures 3(c) and (d) also indicate an improvement for LP mobiles operating under battery level adaptation. For the system that does not prioritize between LP and HP mobiles, we see when comparing the appropriate curves in Figures 3(c) and (d) that the delay performance of the two battery classes are relatively similar. The delay at both LP and HP mobiles starts to grow larger as the traffic increases due to buffering. In the system with battery level adaptation, the packet delay as shown in Figure 3(c) for LP users remains low even for high packet arrival rates when compared to the system without battery level adjustments. The difference between the average packet delay of the two systems gets larger for LP mobiles with increasing traffic load. At high traffic load, the battery adaptive system starts to allocate more and more channels to LP mobiles and thereby counters the effect of the large packet arrival rate at their queues. These lower delay measurements for the LP users are seen for both values of r, but in particular they are lower when there fore few number of LP mobiles.

The delay measurements for the HP users in Figure 3(d), on

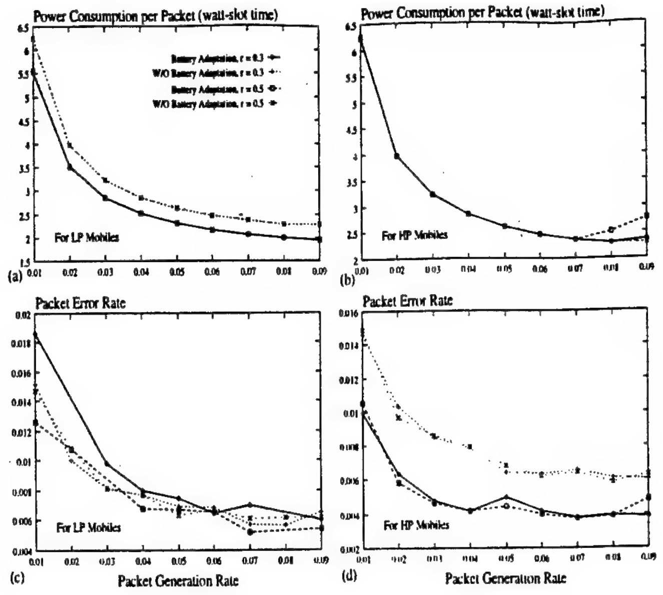


Fig. 4. Power Consumption per Packet for (a) low-power mobiles and for (b) high-power mobiles, and Packet Error Rate for (c) low-power mobiles and for (d) high-power mobiles. r is the ratio of low-power mobiles to the total number of mobiles in the system.

the other hand, indicate a different result. For all traffic loads, we see that in the system with the battery adaptation there is a greater delay on packets originating from HP terminals. Note that at lower traffic rates the average packet delay of the adjusted system is closer to that of the system without battery adaptation. This difference gets significantly larger at very high traffic rates since at these loads LP users are assigned most of the available low-priority slots resulting in longer delays at the HP terminals.

One way to improve the performance of the HP mobiles in the heavy traffic situations is to develop a scheduling algorithm specifically designed to maintain a minimum throughput and maximum packet delay for HP users. One possibility is using a scheduler that prioritizes high-power users once in every few frames. Other such techniques for improving HP performance during heavy traffic are currently being studied.

Curves in Figures 4(a) and (b) indicate battery power adaptation provides energy efficient transmissions for LP terminals. First, we observe in Figure 4(a) that the total power consumed per transmitted packet for LP users under battery power adjustments is less than that for the other system at both ratios due to lower transmit power levels. The power expended in the system without battery adaptation is similar for both LP and HP mobile due the similarities between LP and HP transmissions in this system. The power consumed per transmitted packet is higher during light traffic loads because terminals spend a significant portion of the time in the idle mode since they have a relative small number of packets to transmit. Since significant power is expended at the stand-by level, the power efficiency at low traffic load is larger than during heavy traffic.

In Figure 4(b), we see that HP users expend almost the same power per packet transmission in both systems for low to moderate traffic load. Since the transmit power level is the same for the two systems, the power consumed per transmission is thus the nearly the same in both the systems under analysis. This, however, is not the case at heavier traffic loads when HP users with battery adaptation – for both ratios – expend more power per packet than do HP users in the other system.

The results of the adjustments on the packet error rate can be seen in Figures 4(c) and (d). These figures provide a means to show that our algorithms did not degrade the desired error rate as compared to the system without the power control adjustments. As we see in Figure 4(c), the error rate, measured as the ratio of the number of unsuccessful packet transmissions to the the total number of packets transmitted is similar for LP users in the two systems. Again from Figures 4(c) and (d), we observe that the error-rate performance for the system without battery adaptation is almost the same for both ratios and both battery power levels. This is because the both LP and HP users in this system were given the same transmit power levels, target SIR, and number of simultaneous transmitters. The packet error rate for a larger number of LP mobiles is lower since in this scenario a larger number of packets are transmitted and a greater number of slots are assigned to LP terminals. This increase in the transmission slots helps reduce the error rate. The packet error rates for HP mobiles in the battery adaptive system are lower for both ratios. Since HP mobiles transmit less often in the adjusted system, the total number of transmitted packets reduces and so does the packet error rate.

Using our proposed algorithms, we see that terminals low on battery power can achieve greater throughput and reduce their packet delay. Consequently, this improves communications capability before complete battery loss. The power efficiency helps extend battery usage over time. Finally, as shown by the packet error rate measurements, these benefits are gained without sacrificing the QoS criterion on the target error rate.

5 Summary

The paper focused on techniques to help improve the transmission capabilities of low-power mobiles in a hybrid CDMA/TDMA system. The time-division properties of the system were exploited to schedule transmission times to users based not only on their traffic priority but also terminal battery status. The clustering phenomena that resulted from the proposed scheduling modifications was then used as the basis to modify CDMA power control algorithms. The power control adjustments were geared to adapt to the battery status of each cluster and hence of the serving mobiles. A discrete-event simulation was performed and its results presented. The analysis demonstrated a performance enhancement for low-power terminals operating under our proposed algorithms.

References

[1] J.-C. Chen, K. Sivalingam, P. Agrawal, and S. Kishore, "A Comparison of MAC Protocols for Wireless Local Networks Based on Battery Power Consumption," in *Proc. IEEE INFOCOM*, (San Francisco, CA), pp. 150–157, Apr. 1998.

[2] M. Zorzi and R. R. Rao, "Error control and energy consumption in communications for nomadic computing," *IEEE Transactions on Computers*, vol. 46, pp. 279–289, Mar. 1997.

[3] P. Lettieri, C. Fragouli, and M. B. Srivastava, "Low power error control for wireless links," in *Proc. ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, (Budapest, Hungary), Sept. 1997.

[4] A. Sampath, S. Kumar, and J. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. of IEEE PIMRC*, Sept. 1995.

[5] S. Ramakrishna, "A scheme for throughput maximization in a dual-class CDMA system," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, (San Diego, CA), Oct. 1997.

[6] M. A. Arad and A. Leon-Garcia, "Scheduled CDMA: a hybrid multiple access for wireless ATM networks," in *Proc. of IEEE PIMRC*, pp. 913–917, 1996.

[7] K. M. Sivalingam, J.-C. Chen, and P. Agrawal, "Design and analysis of low-power access protocols for wireless and mobile ATM networks," *ACM/Baltzer Mobile Networks and Applications*, 1998. To Appear.

[8] D. Goodman, *Wireless Personal Communications Systems*. Addison-Wesley Publishing Company, 1997.

[9] K. G. Chen, *Integrated dynamic radio resource management of wireless communication systems*. MS Thesis, Rutgers University, May 1996.

To APPEAR, IEEE PIMRC, SEP. 1998

ON SCHEDULING OF MULTIMEDIA SERVICES IN A LOW-POWER MAC FOR WIRELESS ATM NETWORKS

Jyh-Cheng Chen[1], *Krishna M. Sivalingam*[2], *Prathima Agrawal*[3] *and Raj Acharya*[1]

[1]Department of Electrical & Computer Engineering, State University of New York at Buffalo, Buffalo, NY 14260
[2]School of Electrical Engineering & Computer Science, Washington State University, Pullman, WA 99164
[3]Networked Computing Technology Department, AT&T Labs, Whippany, NJ 07981

ABSTRACT

This paper describes the design and analysis of the scheduling algorithm for EC-MAC (energy conserving medium access control) [1], a low-power medium access control (MAC) protocol for wireless and mobile ATM networks. Based on the structure of EC-MAC and the characteristics of wireless channel, we propose a new algorithm which can deal with the bursty errors and the location-dependent errors. Most scheduling algorithms proposed for either wired or wireless networks were analyzed with homogeneous traffic or multimedia services with simplified traffic models. We analyze our scheduling algorithm with more realistic multimedia traffic models. One of the key goals of the scheduling algorithm is simplicity and fast implementation. Unlike the time-stamp based algorithm, our algorithm does not need to sort the virtual time, thus reducing the complexity of the algorithm significantly.

1. INTRODUCTION

This paper describes the design and analysis of a scheduling algorithm for a low-power medium access control (MAC) protocol for wireless/mobile ATM networks. The design of the protocol – denoted EC-MAC (energy conserving medium access control) – is driven by two major factors. The first factor is that the access protocol should be energy-efficient since the mobiles typically have limited power capacity. The second factor is that the protocol should provide support for multiple traffic types, with appropriate quality-of-service (QoS) levels for each type. In [1, 2], the design and analysis of EC-MAC and the comparison of energy consumption to a number of other protocols have been provided. The core of the protocol, that determines the performance and guarantees the QoS, is the scheduling algorithms associated with the MAC protocols. Such a scheduling algorithm is the focus of this paper. By this scheduling algorithm, we show that EC-MAC, in addition to low energy consumption, can achieve high channel utilization, low packet delay, and meet the QoS requirements for multimedia traffic.

Many queuing and scheduling algorithms have been proposed for conventional wired ATM networks. The framework is that there are queues in switches. The scheduling disciplines then schedule packets in queues accordingly. The schedule disciplines may be as simple as FIFO (first-in-first-out) or round robin, or based on virtual finishing times, such as Virtual Clock [3], SCFQ [4], etc. Previous work for wireless ATM has reported mechanisms for providing the base station with the transmission requests. The scheduling algorithms, however, were not addressed in extensive detail. Recently, research on extending the scheduling algorithms proposed in wired networks to wireless domain has been reported [5]. The major concern addressed here is modification of conventional wired scheduling algorithms to deal with the error-prone wireless channel. However, scheduling with low power in consideration has not been addressed so far. This paper proposes an algorithm that addresses this important issue. Discrete-event simulation with realistic multimedia traffic models is used to obtain the performance results.

2. OVERVIEW OF EC-MAC

The network architecture of EC-MAC is mainly derived from the SWAN network built at Bell Labs [6] – one of the first wireless ATM network testbeds. The access protocol is defined for an infrastructure network with a single base station serving mobiles in its coverage area. The goals of low energy consumption and QoS provision lead us to a protocol which is based on reservation and scheduling strategies.

Transmission in EC-MAC is organized by the base station into frames. Each frame is composed of a fixed number of slots, where each slot equals the basic unit of wireless data transmission. The frame is divided into multiple phases as shown in fig. 1. At the start of each frame, the BS transmits the frame synchronization message (FSM) on the downlink. This message contains framing and synchronization information, the uplink transmission order for reservations, and the number of slots in the new user phase. During the request/update phase, each registered mobile transmits new connection requests and queue status of established queues according to the transmission order received in FSM. The base station then broadcasts the transmission schedule for the data phase using a schedule message. Mobiles receive the broadcast and power on the transmitters and receivers at the appropriate time in date phase.

The next section describes the development and analysis of the scheduling algorithm used for allocating uplink and downlink slots.

3. PROPOSED ALGORITHM

The design of the proposed algorithm is described in the following sections.

The authors can be reached at *jcchen@eng.buffalo.edu*, *krishna@eecs.wsu.edu*, *pa@research.att.com*, and *acharya@eng.buffalo.edu*. Part of the research was supported by Air Force Office of Scientific Research grant F-49620-97-1-0471.

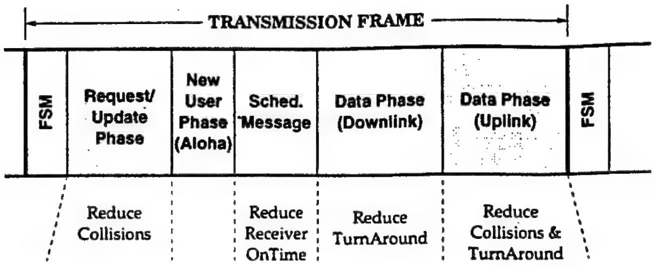


Figure 1: Definition of the different phases in EC-MAC protocol.

3.1. Connection Admission Control (CAC):

After a mobile is admitted to this cell – either locally generated or handed off, it may request bandwidth for several VCs as they are created. The CAC's goal is to maintain QoS for all existent VCs while admitting new VCs.

We use a simple algorithm in which each VC sends the minimum guaranteed number of slots it needs as part of session set up. A counter is used to record the total number of slots which have been admitted. If the counter exceeds the number of slots in uplink data phase after adding this VC's request, the VC is rejected. Otherwise, it is admitted. The counter is incremented by the number of this VC's request slot(s). By this admission control algorithm, the total rate of all admitted VCs is always less than or equal to the maximum capacity in data phase. Therefore, the QoS of existing VCs will not be affected by the newly admitted VC.

3.2. Scheduling

The proposed algorithm performs *coarse-grained scheduling* based on the frame structure of EC-MAC. Although many algorithms have been proposed for conventional wired ATM networks [7], most of them are based on packet-by-packet scheduling which are good for *fine-grained scheduling* only. For example, algorithms based on time-stamp such as Virtual Clock [3] and SCFQ [4] are not applicable for EC-MAC because they need to know the arrival time of each packet. Other type of algorithms such as HOL-EDD [8] might be modified for frame-based scheduling. Since it does not allow a session to be served at different rates at different times, a VBR (variable bit rate) video session cannot improve the delay performance without requesting the peak bandwidth. A *multirate* service algorithm was proposed to address the scheduling of VBR video [9]. However, this algorithm is fine-grained based on time-stamped priority. In addition, it was proposed for high-speed networks where errors are negligible. In wireless domain, algorithms such as that proposed in [5] do not consider the energy consumption factor. In addition, realistic multimedia traffic models are not used to investigate the performance. It is uncertain how well these algorithms will perform for multimedia terminals with limited battery power.

The proposed algorithm is a *priority round robin with dynamic reservation update and error compensation* scheduling. The scheduler is currently defined to handle CBR (constant bit rate, e.g. voice), VBR (variable bit rate, e.g. video), and UBR (unspecified bit rate, e.g. data) traffic. The scheduler gives higher priority to CBR and VBR traffic. These traffic sources can make requests for slot reservations that will be satisfied by the scheduler. UBR traffic, on the other hand, is treated with low priority and without reservation. Within the same traffic type, the different connections are treated using round robin mechanism.

The base station (BS) maintains two tables: *request table* and *allocation table*. The request table maintains the queue size of the virtual circuit of each mobile, the error state of the mobile, the number of requested reservations for CBR and VBR traffic, and the number of credits for UBR traffic. The purpose of the allocation table is to maintain the number of slots scheduled for each VC and each mobile. This table is essentially broadcast as the schedule to the mobiles. Based on this table, the base station allocates contiguous slots within a frame for each mobile.

The BS first allocates slots to CBR VCs which have been currently admitted. Because of the connection admission control (CAC) described above, CBR VCs that belong to mobiles in non-error (good) states are satisfied with their required rates. The CBR VCs are allocated X slots every Y frames, based on the traffic requirements. For instance, with a 12-ms TDMA frame, a 32-Kbps voice source is allocated one 48-byte slot per frame.

For sources with VBR traffic, the base station maintains the number of slots allocated in the previous frame. Let the current request of source i be C_i slots, and the allocation in previous frame be P_i slots. If $C_i < P_i$, C_i slots are allocated, and the remaining $P_i - C_i$ are released. If $C_i > P_i$, P_i slots are allocated in the first round. In the second round, extra slots available are evenly distributed among the VBR sources whose requests have not been fully satisfied in the first round.

Since there is correlation in a VBR video source, the reservation in current frame period represents the prediction for next frame. By the adjustment, the bandwidth allocation in each frame is different depending on the current traffic load and the number of packets generated by VBR sources. The reservation, hence, is updated dynamically in each frame for VBR traffic.

The BS then schedules UBR traffic after the scheduling of CBR and VBR. If the mobile is in error state, the base station adds credit(s) in the corresponding entry in request table. Otherwise, the base station either schedules slot(s) to this VC or schedules the aggregate credits this VC has until there is no more slot available. The reason for this credit is to ensure long-term fairness. This credit adjustment scheme is not applied to voice and video traffic since late packets will be dropped rather than be played back in such applications.

3.3. Contiguous bandwidth allocation:

The scheduling algorithm updates the corresponding entry in allocation table as described above for each VC request. The allocation table can be implemented as a two-dimensional array with one dimension for each mobile and the other dimension for each VC of this mobile. The base station broadcasts the slot id and the number of slots for each VC by looking at the entry of each mobile. The pseudo code is listed in fig. 2. By this allocation table, each mobile listens to all schedule information meant for it contiguously. It also gets slot allocation in data phase contiguously for all different traffic types although the scheduling is done on the basis of traffic type. Therefore, mobiles only need to turn on transmitter/receiver once each during schedule phase and data phase. Please note the total number of allocated slots in allocation table is less or equal to the total slots in data phase. This has been checked in the scheduling algorithm described above. By the algorithm in fig. 2 and the allocation table, the slot allocation is announced on a frame-basis rather than on a slot-basis. Mobiles also only need to turn on the

```
BROADCAST_SCHED ()
/* Broadcast sched. msg. based on allocation table */
/*
$N_m$:              Total number of mobiles in the system;
$N_{vc}$:           Maximum number of VCs in each mobile;
$slod\_id$:         Beginning slot in data phase for each VC;
$Sched$ [ ]:        One-dimension array for each sched. beacon;
$Allocate$ [ ] [ ]: Two-dimension array of the allocation table ;
*/

$index = slot\_id = 0$;
/* for each mobile in the array */
for $(i = 0;\ i < N_m;\ i++)$
    /* for each VC in the mobile */
    for $(j = 0;\ j < N_{vc};\ j++)$
        /* if the entry is not zero */
        if $(Allocate[i][j]\ != \ 0)$ {
            /* announce the slot allocation in the
               schedule beacon */
            $Sched[index].macid.id = i$;
            $Sched[index].macid.vc = j$;
            $Sched[index].slot\_id = slot\_id$;
                /* beginning slot in data phase */
            $Sched[index].slot\_num = Allocate[i][j]$;
                /* number of allocated slot(s) */

            $slot\_id += Allocate[i][j]$;
                /* increment the slot id */
            $index++$;     /* next schedule beacon */
        }
```

Figure 2: Contiguous bandwidth allocation.

transceiver once for all different types of packets.

3.4. Dealing with Errors

This section describes how the scheduling algorithm deals with channel errors. At a time, only some of the mobiles may be capable to communicate with the base station – the others might be in error state. Since a mobile may encounter errors during any phase of the time frame, we discuss them individually as follows.

1. If a mobile is in error state during base station frame synchronization message (FSM) reception, it will not receive its transmission order. Thus, it will not send the request in the uplink of reservation phase, and the BS will mark the mobile as in error state. The scheduling algorithm might assign credits to the mobile depending on the traffic type.

In case the mobile changes to good state any time after this phase, the mobile will not be able to transmit in the subsequent data phase. It could decide to receive broadcast packets.

2. If errors happen during the uplink of request/update phase, the BS will mark the mobile as in error state because it did not receive the transmission request. When the mobile sends request in the subsequent request/update phase, BS will mark the mobile as in good state. The situation is similar to the one above.

3. If errors happen while a mobile is receiving the schedule message, bandwidth that has been scheduled to this mobile will not be utilized. This loss is limited to only one data phase which is typically smaller than the average burst error length. The BS will mark the mobile as in error state when it does not receive this mobile's data during the scheduled uplink slots. The BS will mark the mobile as in good state when it receives the requests from mobiles in request/update phase again.

4. If errors happen during the downlink data phase or the mobile does not turn on receiver because of missed schedule message, the BS will hold packets until the corresponding mobile returns back to good state. Mobiles can acknowledge the packets they receive when they send requests in next request/update phase. Thus, the BS can know whether mobiles have received the downlink packets or not. The BS deletes packets from queues only after it receives acknowledgments.

5. If errors happen during the uplink of data phase, the BS will not receive the packets sent from the mobiles in error state. The BS acknowledges the packets it received in the next FSM. Mobiles delete packets from queues only after receiving acknowledgments or the deadline of real-time packets is expired. The BS will know the actual queue size of each VC and reschedule the packets when it receives the requests from mobiles in uplink request/update phase again.

This section described the mechanisms defined in EC-MAC to handle channel errors during the various phases. The following section provides a simulation based performance analysis.

4. PERFORMANCE ANALYSIS

The following sections describe source traffic models and simulation results for the algorithm described above using realistic source traffic models for video, voice, and data services.

4.1. Source Models

The simulation results presented here consider three types of traffic – each for CBR, VBR, and UBR category. Voice is modeled as a two-phase process with talkspurts and silent gaps [10]. Typically, such modeling classifies voice as VBR. We consider that the voice source generates a continuous bit-stream during talkspurts and is therefore classified as a CBR source in our scheduling. Video is considered as an example of a VBR source with variable number of cells per frame. Data generated by applications such as ftp, http and email is considered as an example of UBR traffic.

In simulation, each mobile terminal is capable of generating three different types of traffic: data, voice, and video. An idle mobile generates new voice calls and video calls with rates of λ_s, and λ_v, respectively. Data traffic is modeled as self-similar traffic with Hurst parameter of 0.9 (described below). The following paragraphs present the simulation models for data, voice, video, and error, respectively.

Data Model: Recently, extensive studies show that data traffic is self-similar in nature, and the traditional Poisson process cannot capture this fractal-like behavior [11]. Long-range dependent traffic (fractional Gaussian noise) can be obtained by the superposition of many ON/OFF sources in which the ON and OFF periods have a Pareto type distribution with infinite variance [11]. In simulation, we use the strictly alternating ON/OFF sources with the same α-value for the Pareto distribution. The α value equals 1.2 which corresponds to the estimated *Hurst parameter*, the index of self-similarity, of $H = 0.9$ [11].

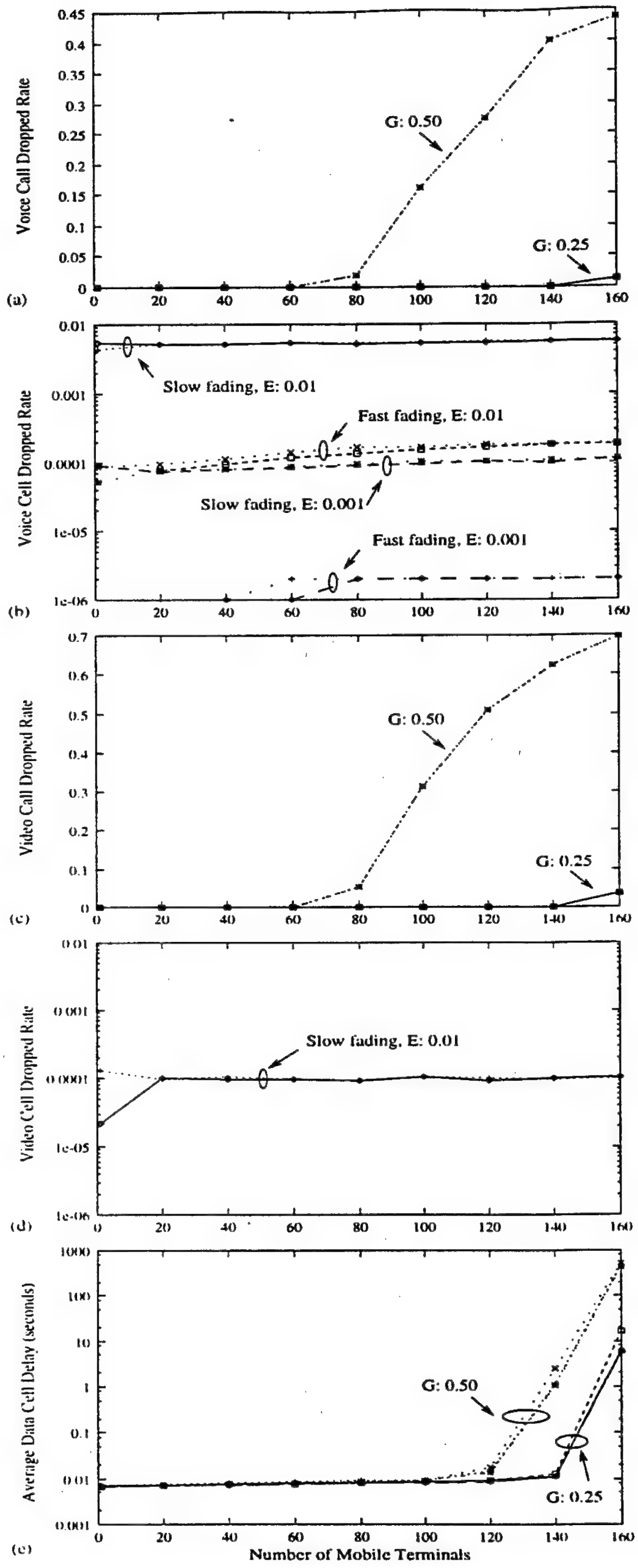


Figure 3: Performance results, where E is the packet error rate, and G is the traffic load in each mobile.

Voice Model: A voice source is modeled as a two-state Markov process representing a source with a *slow speech activity detector* (SAD) [10]. Measured values for talkspurts and silence are 1.00 sec and 1.35 sec [10], each with exponential distribution. This results in an average of 36% talkspurts and 64% silence gaps for each voice conversation. A voice cell is dropped if not transmitted after 36 ms. When a new voice cell arrives at a full queue, the *first* cell in the voice queue will be dropped.

Video Model: H.263 video targets the transmission of video telephony at data rates less than 64 Kbps which makes it suitable for wireless communications. In simulation, we used the real trace data from several H.263 video sources [12]. Each video runs for around 30 seconds. The frame rate is 25 fps for all videos. For a TDMA frame of length 12 ms (as used in the simulation), the mean number of video packets is around 1 ATM cell per TDMA frame and the maximum is 21 ATM cells per TDMA frame. In the simulation, we assume that the length of a video session is exponentially distributed with mean time of 5 minutes. This is achieved by randomly selecting different videos (since each video trace only lasts above 30 seconds).

Error Model: In wireless networks, errors are bursty and location-dependent. Models for a Rayleigh fading channel have been studied in [13]. Here, it has been shown that a first-order Markov model is an adequate approximation for a Rayleigh fading channel. In our studies, we use the slow and fast fading models proposed in [13]. The values of normalized Doppler bandwidth for slow fading and fast fading are 0.01 and 0.64, which correspond to users with moving speed about 1.5 km/h and 100 km/h, respectively. Two fading margins are considered: 29.9978 dB and 19.9782 dB, which correspond to packet error rates of 0.001 and 0.01, respectively [14].

4.2. Simulation Results

The numerical results presented here study the maximum number of mobiles that can be accommodated with the desired QoS for voice, data, and video traffic. A channel rate of 10 Mbps has been considered. Each uplink and downlink in data phase is around 4.7 Mbps. The figures are plotted with offered load of 25% (G = 0.25) and 50% (G = 0.50) per mobile. Data traffic is modeled as self-similar traffic with Hurst parameter of 0.9 [11]. When load is 50%, the inter-arrival times of voice calls ($1/\lambda_s$) and video calls ($1/\lambda_v$) are 180 sec and 300 sec, respectively. The average length of a voice call is 3 minutes, so the voice traffic load is $180/(180 + 180) =$ 50%. The average length of a video call is 5 minutes, so the video traffic load is $300/(300 + 300) = 50\%$ also. The *packet error rates* are 10^{-2} and 10^{-3} with fast fading and slow fading.

Voice-call dropped rate is considered first in fig. 3(a), where same traffic load with different error rates and fading models leads to same results. Therefore, only two lines are visible in fig. 3(a). As expected, less voice calls are dropped if less mobiles contend in the system. With the same number of mobiles, traffic load of 50% leads to higher dropped rate than traffic load of 25%. Fig. 3(a) also shows that the major factor for call dropped rate is traffic load rather than the fading or error rate because the call dropped rate depends mainly on CAC. When the load is 50% and the number of mobiles is greater than 80, the voice call dropped rate increases rapidly. For the load of 25%, voice call dropped rate is acceptable even when the number of mobiles is 160 regardless of the error rate.

Once a voice call is admitted, fig. 3(b) indicates that the voice-cell dropped rate is almost independent of traffic load. In fig. 3(b), each set of fading and error rate is simulated for two traffic loads: 0.25 and 0.5. The cell dropped rate is less than the error rate for slow fading regardless of the number of mobiles and the load offered by each mobile. The cell dropped rate is much less than the error rate for fast fading. The difference in fast fading and slow fading is that the slow fading is more bursty than fast fading [14]. This leads to a shorter error period when error happens in fast fading. Each voice cell can tolerate $36ms$ delay. Longer error period may cause more expired voice cells. Hence, slow fading has higher dropped rate. Fig. 3(b) shows that the cell dropped rate increases slightly for fast fading when the number of mobiles increases.

The CAC algorithm restricts the number of connections to maintain the QoS of admitted connections. The CAC and scheduler cooperate with each other like this: CAC deals with traffic load and scheduler deals with QoS requirements. Although cells are still dropped, that is the *de facto* nature of wireless channel due to errors. Fig. 3 shows that the call dropped rate which depends on traffic load is determined by CAC, while cell dropped rate, depended on fading and error rate, is determined mainly by scheduler.

Figs. 3(c) and (d) examine the video-call and video-cell drop rates. As discussed above, call dropped rate is determined mainly by the offered traffic load. With CAC, a video call sends the minimum guaranteed rate it needs. Based on this information, CAC decides to admit or reject this call. If a video call requests the maximum rate it needs in CAC, there will be no dropped cell ideally. However, it is wasteful to decide on admission control based on maximum bandwidth requirement. If it requests a mean rate in CAC, many other sessions can be admitted but the cell dropped rate may be unacceptable. For a H.263 video with mean rate of 1 ATM cells per TDMA frame and peak rate of 21 cells per TDMA frame, figs. 3(c) and (d) show the results when each video sets 2 cells as the minimum guaranteed rate. Although the request rate set in CAC is still much less than the peak video rate, fig 3(d) indicates that the video-cell dropped rate is very small even with 0.01 error rate for slow fading. All others almost have 0 video-cell dropped rate in fig. 3(d). This indicates that our *dynamic reservation update* scheme can get a good multiplexing gain. Fig. 3(c) also shows our algorithm can support 80 video sessions when load equals 0.5. More than 160 video sessions can be accepted when load equals 0.25 if the required call dropped rate is set to 1%.

Fig. 3(e) compares the data cell delay. Data traffic is transmitted when there are no other voice or video traffic pending. Although data is with lower priority and without any reservation, it still gets chances to transmit when some voice or video sessions are in error state, or when VBR video sessions generate less traffic. As expected, the data cell delay increases when the number of mobiles increases. Fig. 3(e) shows that the higher load generally has higher data delay. Channel fading and error rate, however, will not affect data delay too much. This is because the scheduling algorithm credits the error mobiles after they change back to good state.

5. CONCLUSION

This paper describes a scheduling algorithm for EC-MAC, a low-power access protocol for wireless and mobile ATM networks. The goals of the access protocol are to conserve battery power, to support multiple traffic classes, and to provide different levels of service quality for bandwidth allocation. The proposed algorithm is a *priority round robin with dynamic reservation update and error compensation* scheduling. Performance analysis based on discrete simulation is presented. The analysis studies various quality-of-service parameters with varying number of mobiles in a cell. Low-power operation current is done by contiguous bandwidth allocation and by cooperating with EC-MAC [1, 2]. Future work will include more power adaptation by prioritizing low-power and high-lower mobiles in scheduling.

REFERENCES

[1] K. M. Sivalingam, J.-C. Chen, and P. Agrawal, "Design and analysis of low-power access protocols for wireless and mobile ATM networks," *ACM/Baltzer Mobile Networks and Applications*, 1998. To Appear.

[2] J.-C. Chen, K. M. Sivalingam, P. Agrawal, and S. Kishore, "A comparison of MAC protocols for wireless local networks based on battery power consumption," in *Proc. IEEE INFOCOM*, (San Francisco, CA), pp. 150–157, Apr. 1998.

[3] L. Zhang, "VirtualClock: a new traffic control algorithm for packet switching networks," *ACM Transactions on Computer Systems*, vol. 9, pp. 101–124, May 1991.

[4] S. Golestani, "A self-clocked fair queueing scheme for broadband applications," in *Proc. IEEE INFOCOM*, (Toronto, Ont., Canada), pp. 636–646, June 1994.

[5] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," in *Proc. ACM SIGCOMM*, (Palais des Festivals, Cannes, France), Sept. 1997.

[6] P. Agrawal, E. Hyden, P. Krzyzanowski, P. Mishra, M. B. Srivastava, and J. A. Trotter, "SWAN: A mobile multimedia wireless network," *IEEE Personal Communications*, vol. 3, pp. 18–33, Apr. 1996.

[7] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE*, vol. 83, Oct. 1995.

[8] M. Vishnu and J. W. Mark, "HOL-EDD: A flexible service scheduling scheme for ATM networks," in *Proc. IEEE INFOCOM*, (San Francisco, CA), pp. 647–654, Apr. 1996.

[9] D. Saha, S. Mukherjee, and S. K. Tripathi, "Multirate scheduling of VBR video traffic in ATM networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1132–1147, Aug. 1997.

[10] D. J. Goodman and S. X. Wei, "Efficiency of packet reservation multiple access," *IEEE Transactions on Vehicular Technology*, vol. 40, pp. 170–176, Feb. 1991.

[11] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 71–86, Feb. 1997.

[12] "Digital video coding at Telenor R&D." http://www.fou.telenor.no/brukere/DVC/.

[13] M. Zorzi, R. Rao, and L. B. Milstein, "On the accuracy of a first-order Markov model for data transmission on fading channels," in *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, pp. 211–215, Nov. 1995.

[14] A. Chockalingam, M. Zorzi, and R. Rao, "Performance of TCP on wireless fading links with memory," in *Proc. IEEE International Conference on Communications (ICC)*, (Atlanta GA), June 1998.

IN IEEE PIMRC '98, SEPT. 1998

An Experimental Architecture for providing QoS guarantees in Mobile Networks using RSVP

Indu Mahadevan[†] and Krishna M. Sivalingam[‡*]
School of Electrical Engineering and Computer Science
Washington State University
Pullman, WA 99164

ABSTRACT

Efforts are underway to enhance the Internet with Quality of Service (QoS) capabilities for transporting real-time data. The ReSerVation Protocol (RSVP) provides a signaling mechanism for end-to-end QoS negotiation. The issue of wireless networks and mobile hosts being able to support applications that require QoS has become very significant. Reservation of resources and the maintenance of QoS for the mobile as it moves from one region to another creates a new set of challenges. In our paper, we describe an architecture where a modified RSVP protocol helps provide QoS support for mobile hosts. The modified RSVP protocol has been implemented in an experimental wireless and mobile testbed to study the feasibility of our approach.

I. INTRODUCTION

Future wireless and mobile communication networks will be expected to provide resource allocation for the various classes of applications that require Quality of Service (QoS) support. A big drop in service quality when a call hand-off is made as the mobile moves from one region to another may not be acceptable for these applications. It is required to maintain the QoS of these applications, in the presence of user mobility with the use of resource reservation. ReSerVation Protocol (RSVP) [1, 2] is a network management setup protocol designed to help share resource reservations among participating applications. Currently, RSVP is designed to operate in wired networks. In this paper, we will describe the design and architecture of a modified RSVP protocol to guarantee resource reservations to mobile wireless hosts.

In our architecture, a mobile in a region is served by a base station which is connected to the wired network. Resource reservations are made using RSVP between the base station and the mobile. To make sure that a mobile has reservations guaranteed as it moves from one region to another, base stations make reservations with other base stations in all the neighboring regions. These reservations will remain "passive" [3]. That is, the resources may be used by other mobiles until it is needed for this particular mobile. This ensures that resources are not needlessly tied up for potential incoming mobile hosts.

An architecture for using RSVP for Integrated Services Packet Network with mobile networks has been described in [3]. A passive reservation mechanism is suggested in [3] as described above. However, the architecture requires a mobile to know all the subnets it will be visiting. The mobile obtains the identity of the proxy agents, which help with mobile RSVP in all the subnets, using a proxy discovery protocol. The mobile instructs the proxy agent in the region it is currently located to make passive reservations with all the proxy agents in all other regions. Four additional messages are used in addition to the messages already present in RSVP. The drawback of this architecture is that it assumes that a mobile knows the addresses of all the subnets it is going to move into and which is not always possible. It also places a burden of finding the proxy agents in all these subnets on the mobile.

In this paper we will discuss our architecture which does not make these assumptions and show how we can enable RSVP to work with mobile networks so that QoS can be maintained. The work addresses messages conveying resource requests and scheduling at the Base Station (BS) to accommodate resource requests. Class Based Queueing (CBQ) [4] is used as the underlying packet classifying and scheduling mechanism in our architecture. The experimental testbed developed based on our architecture, consists of base stations and laptops equipped with WaveLAN [5] cards and WavePOINTs that help with roaming support. We use RSVP which is modified to identify "active" and "passive" reservations.

The goals and results from our paper are summarized here. We mention the changes made to RSVP and CBQ to experimentally substantiate the architecture. We have considered QoS parameters specific to the mobile environment. These parameters – loss profiles, probability of seamless communication and rate reduction factor are described in detail in section IV. We have modified a few applications including a traffic generator, a video decoder and a benchmarking program to use RSVP for resource reservation. The results show the establishment of a passive reservation with a neighboring BS, that is later converted to active reservation after hand-off and the incorporation of loss profiles to influence CBQ's packet drop mechanism.

[*] Corresponding Author. Part of the research was supported by Air Force Office of Scientific Research grant F-49620-97-1-0471. E-mail: [†]imahadev@eecs.wsu.edu. [‡]krishna@eecs.wsu.edu.

The paper is organized as follows. Section II gives the outline and features of our architecture. Section III gives an overview of RSVP and CBQ discusses using RSVP for this problem. We look at QoS specifications in section IV. Section V discusses the experimental testbed followed by results in section VI. Conclusions follow in section VII.

II. THE NETWORK ARCHITECTURE

In this section we introduce our architecture which uses RSVP to reserve resources in a mobile environment.

This paper assumes a microcellular network architecture, with a geographical region divided into cells. Each cell has a Base Station (BS) serving all mobiles within its coverage region and connected to the wired network. When a mobile moves to another cell, it is handed off to the base station serving that cell.

In our network architecture, we use RSVP along with Class Based Queueing (CBQ) to reserve resources in the network. The sender or receiver of an application can be a mobile, the base station or any host on a wired network. In the discussion here, we emphasize on the communication between the base station and the mobile. Reserving resources is required to make sure an application gets the required QoS. The problem with wireless mobile networks is to make sure we maintain the QoS as the mobile moves from one cell to another. This means we need to make sure that the mobile has some form of resource reservation anywhere it goes.

One way of guaranteeing QoS as a mobile moves from one cell to another is to reserve resources with all the base stations in the neighboring cells because the mobile might move into one of them. This would be a waste of limited wireless resources. Alternatively, we can make reservations that can be used by other mobiles in the cells till the mobile moves into that cell. These reservations made in the neighboring cells will be "passive" and can be used for other applications till the mobile actively starts using them [3]. An architecture to make such resource reservations using RSVP is discussed below.

A. Scenario with Reservations for Mobiles

Fig. 1 aids the discussion in this section. The cells are denoted by A, B, C, etc. BS represents the base station and M represents the mobile. The solid line represents an active reservation while the dotted line indicates a passive reservation. We assume that a base station knows the addresses of the base stations in all the neighboring cells.

In a wireless environment we need to distinguish between the two kinds of reservations that need to be made: a) between the sender and various base stations in neighboring cells in the wired network, and b) between the base station and the mobile in the wireless region. In our example, the mobile M is initially in cell A. BSa is the base station in this cell and resource reservations must be made between BSa and M. When the mobile moves, it could move into any of the other six cells. At this point we will need to make two kinds of resource reservations that will remain "passive": a) one between the "current" base station BSa and all the other base stations namely BSb, BSc, BSd, BSe, BSf and BSg, and b) the other where the base stations BSb, BSc etc. make a passive resource reservation on their wireless interfaces to accommodate a mobile that may enter their cells. In the example provided above, the base stations are assumed as the end points of an application. This need not be the case. If another host is the end-point, a reservation needs to be made between that host and the base station in the wired domain using regular RSVP requests.

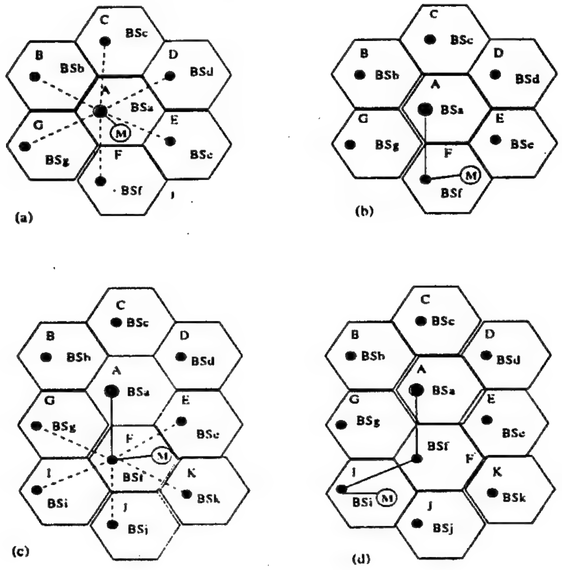


Figure 1: Overview of a Mobile Environment

In fig. 1(b) we see that the mobile M has moved into cell F. At this point the resources that were labeled passive in the wired environment between BSa and BSf are made active and the resources made on the wireless interface of BSf is also activated and is used for communication between BSf and M. All other resources reserved passively can be deleted now. This scenario further continues as shown in figures 1(c) and (d). At some point, we could make a re-routing decision so that the sender BS directly connects to the BS in a cell where a mobile is currently located instead of going through all the intermediate BSs. This decision will be a trade-off between the re-routing cost and the cost of reserving resources in all the intermediate BSs.

B. Features of Our Architecture

One of the major features in our architecture that we would like to emphasize is that we do not assume that a mobile knows the cells it will visit. This kind of information may not be available to the mobile all the time. Our architecture only requires that a base station knows the addresses of the base stations in all the neighboring cells. This is a more reasonable assumption to make because it is easier to get this information on a wired network and this information does not change frequently. Also to be noted is the fact that a base station needs to keeps track of a fixed number (six) of addresses only.

In addition, our architecture introduces some QoS parameters into RSVP which are specific to the wireless environment. We also substantiate the feasibility of our architecture by experimental results. In the next section we provide an overview of RSVP and CBQ and give details of how it is used in our architecture.

III. USING RSVP AND CBQ IN OUR ARCHITECTURE

In this section we will discuss how RSVP and CBQ can be used to reserve resources in a mobile environment. An overview of RSVP and CBQ is followed by a detailed example of how RSVP and CBQ can be used in our proposed architecture.

A. Overview of RSVP and CBQ

The RSVP protocol [1, 2] is a network management setup protocol designed to share resources among participating applications. After a high-level dialogue, the initiator of a flow that wants to use RSVP generates PATH messages to each accepting receiver. The PATH message specifies the upper limits of the flow expected and the message is routed to the receiver(s) using any routing algorithm that is available. The receiving host(s) will then make resource reservation requests using RESV messages along the reverse path to the sender. If at any point along the path the request cannot be supported, that request is blocked. Otherwise, this request is merged with other requests to the same sender in order to better share the bandwidth. Each node (host or routers) that is capable of QoS control needs a packet scheduler and classifier which is handled by Class Based Queueing (CBQ) [4] mechanisms. CBQ consists of a classifier that classifies packets into one of a set of pre-defined classes, an estimator that estimates bandwidth usage of each class and a packet scheduler that selects the next class to send a packet. RSVP is independent of the underlying scheduling mechanism and can be aided by any mechanism like CBQ for QoS control.

B. Using RSVP and CBQ to reserve resources

Using RSVP and CBQ to reserve resources is discussed using fig. 2 with two neighboring cells – Cell1 and Cell2. We assume, for simplicity of discussion, that a base station (BS) is the sender of an application. We use the RSVP protocol, modified to recognize passive and active reservations. The mobile is currently in Cell1 in which the sender/BS resides. The BS sends PATH messages to the mobile and since the reservations are going to be used it is indicated as an active reservation denoted in Fig. 2(a) as [1]. The mobile responds with a RESV message (active one) if it can accept the call. After this point, the sender of the application has made sure that resources are reserved for this application. The BS now has to make passive reservations with all the BSs' in the neighboring cells. The traffic specifications of this passive reservation are the same as those used by the active reservation. In fig. 2 the "current" BS 1 sends PATH messages denoting it is a passive reservation to BS 2 (shown as [3] in fig.2) and the BS 2 will send a "passive" RESV message to BS 1 (shown as [4] in fig. 2). BS 2 also needs to make reservations on the wireless interface (shown as [5] in fig. 2) for the wireless link which the mobile may use.

Once the mobile has moved to the neighboring cell, both the mobile and the BS will know that a hand-off has been made. At this point, PATH and RESV messages denoting an active reservation

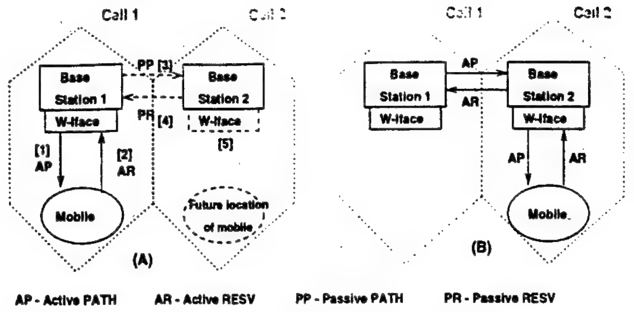


Figure 2: RSVP Messages for Reservation

are exchanged between BS 1 and BS 2 and between BS 2 and the mobile. Reservation between BS 1 and the mobile can be deleted by sending TEAR message or by just letting the the reservation to be deleted by the lack of refreshing PATH and RESV messages. In the discussion above when we say that we have a reservation, we mean that RSVP has done the end-to-end QoS negotiation and the data sent from an application will now use CBQ for packet scheduling and classification. CBQ will now make forwarding decisions on the outgoing interface which will help in achieving the promised QoS on the particular link-layer medium used by that interface. In the shared wireless environment like ours, CBQ makes decisions on the outgoing interface for the wireless downlink protocol between the BS and the mobile.

The RSVP messages must signal some mobility-specific QoS parameters also. A discussion on the QoS parameters specific to a mobile environment is entailed below.

IV. QoS PARAMETERS FOR MOBILITY

In this section we take a general look at the various QoS parameters that need to be taken care of in a mobile environment and how they can be specified and handled. QoS parameters for real-time services include parameters like packet delay, packet loss rate, delay jitter and minimum and maximum bandwidth which are specified and handled by Integrated Service classes supported by RSVP and CBQ.

Integrated Services offers QoS based on three service classes: *Guaranteed Service* which provides a firm bound on data throughput and delay along the path, *Controlled Load Service* where probabilistic promises are made to provide some service and *Best Effort Service* with no performance guarantees. Applications using Integrated Services Specification are characterized by parameters like *token bucket rate, token bucket depth, peak data rate, minimum policed unit and the maximum packet size.*

Apart from the above mentioned QoS parameters, there are certain parameters that will help deal with problems which are unique to a mobile computing environment [6]. One of the parameters is the *loss profiles* which gives the applications an opportunity to choose whether a bursty loss or a distributed loss is preferred in case of an overloaded situation. This will be based on the nature of the application. For example, an audio stream may prefer distributed loss because the output would still be tangible whereas a video stream may prefer a bursty loss because it would probably appear as a flicker in the output. The second QoS parameter is the *probability of seamless communication* which defines the nature of breaks that can be allowed in the service. The third parameter we introduce, *rate reduction factor*, deals with the proposed passive reservations. Since a BS has to make passive reservations in advance with all the neighboring BSs, there is a possibility that some of these reservations would be turned down because of lack of resources. This parameter denotes a factor by which the original resource request can be reduced in case a reservation does not go through. Such a mechanism will ensure that at least some resources are reserved.

We now consider how these QoS parameters can be handled. Loss profiles can be incorporated by changing the way packet dropping is done at the queues by CBQ. CBQ can drop every "1 in n" packets for a distributed loss and drop few packets is a row for a bursty loss. Seamless communication can be obtained by making sure that when a mobile moves from one cell to another, the packets that could be delayed or lost during the call hand-off are already buffered in the new cell. This can be done by multicasting some data to the neighboring cells ahead of time [7]. The rate reduction factor can be implemented by re-negotiating using RSVP PATH and RESV messages.

V. THE EXPERIMENTAL TESTBED

We have set up an experimental testbed to implement and test our proposed architecture. The testbed has two Pentium systems that operate as base stations. Each base station is equipped with an Ethernet card and a 2.4GHz WaveLAN ISA card. The base stations are in adjacent cells and the different Network Identifiers (NwID) of the WaveLAN cards and WavePOINTs [5] in these cells identify the cells. WavePOINTs identify the systems in a particular cell with the beacon signals where the beacon contains the NwID. The testbed also has two mobiles which are equipped with 2.4GHz PCMCIA WaveLAN cards. The FreeBSD WaveLAN driver for PCMCIA cards supports roaming [8] and is used in conjunction with WavePOINT. The theoretical bandwidth of a WaveLAN card is 2Mbps. The base station and the mobiles run FreeBSD 2.2.2. Our testbed uses RSVP code version 4.2a2 [9] and alternate queueing package version 0.4.2 [10].

In the experimental setup, the base station is physically different from the WavePOINT. A WavePOINT produces beacons and helps the mobile identify a particular cell.[1] Each cell is identified by a particular NwID. The mobile uses the signal strength of the beacon packets from WavePOINT to decide which cell it is in. Only wireless interfaces with the same NwID can communicate with each other. When a mobile moves into a cell the mobile "acquires" the NwID of the WavePOINT and is thus able to communicate with all the wireless interfaces in that cell.

When a mobile moves into a new cell only the WavePOINT and the WaveLAN driver of the mobile are able to identify this move. We will need to convey this information to the "previous" and "current" base stations. Two solutions are proposed and tested:

Base station broadcast method: In this method, the base station periodically broadcasts a message on a known UDP port. This helps in identifying if a mobile has moved into the region.

Mobile broadcast method: In this method, a mobile periodically checks if the NwID of its WaveLAN driver has changed. A change indicates that the mobile is in a new cell.

Apart from Integrated Services parameter, RSVP has been modified to carry the mobility parameters and also a parameter to indicate if a reservation is passive. The loss profiles parameter is conveyed by the application to CBQ through RSVP PATH and RESV messages. CBQ has been modified to introduce bursty or distributed loss based on the loss profiles parameter.

The FreeBSD roaming driver from CMU [8] was modified for use with our testbed. A few applications were modified/written to use RSVP API (RAPI) interface of RSVP to help these applications request resources. The applications included a traffic generator program, a video decoder and a benchmarking program.

VI. EXPERIMENTAL RESULTS

In this section, we describe the experiments conducted to show that our reservation scheme works. We also show how the loss profiles scheme in CBQ works in our architecture. The experiments were monitored with the help of tele traffic trapper (ttt) which comes with FreeBSD, and the CBQ monitor that comes with the ALTQ package [10] in FreeBSD. (The legends from the output of ttt and CBQ monitor were modified to increase clarity. This does not affect the results).

A. Experiments on Passive Reservation

The experimental set up consists of three users (User1, User2 and Mobile) which have reservations for 30%, 22% and 35% of the bandwidth respectively (fig. 3). Currently the mobile is not in the region of the BS where the traffic is being monitored and hence does not contribute to the total traffic in the system. Initially User1 and User2 are sending data that was reserved for them. After a while, User1 starts sending more data than what was alloted to it. User1 is provided the necessary bandwidth because the reservation made for the Mobile is passive and is not yet used by the Mobile. We see that unused reserved bandwidth is not wasted if another application needs it. After a hand-off occurs, the mobile moves into the region and the passive reservation is made active. Now the Mobile starts sending data and so the bandwidth of User1 reduces to its reserved rate (about 30%). User2 is conforming to its reserved rate and hence is not affected. From this experiment we see that (i) passive bandwidth is used by other applications if needed and (ii) when the reservation becomes active, the application that was using this bandwidth has to relinquish it.

B. Experiments with the Loss Profiles Parameter

The loss profiles parameter in a mobile environment defines how the packet drop should be done in lossy situations. An application specifies what kind of loss parameter it wants in the PATH and RESV messages of RSVP. This loss profiles parameter is conveyed to CBQ, which has been modified to use the loss profiles

[1] WavePOINT can also act as a bridge to the Ethernet domain but in our testbed it is used only as a beacon generator.

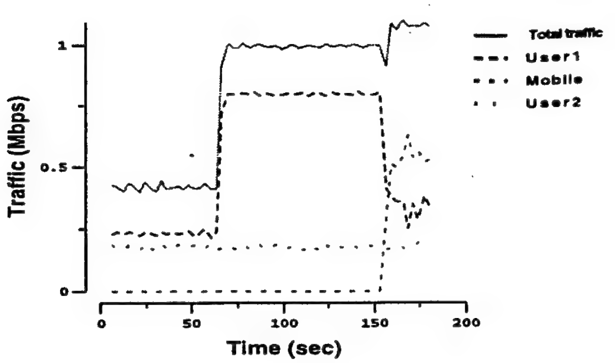


Figure 3: Usage of passive reservation: User1, User2 and Mobile have reserved 30%, 22% and 35% of bandwidth. Initially User1 and User2 send data according to their reservation requests. After a while User1 exceeds its reservation but is accommodated because the Mobile is not yet using its passively reserved bandwidth. After a hand-off process, the Mobile starts using its reserved bandwidth. As a result the excess resources granted to User1 is taken back and User1 reverts to using its original alloted bandwidth.

parameter. Based on the loss profiles parameter the packets are dropped in bursts or in a distributed manner. The default way to drop packets is the tail drop method used by CBQ.

The experimental setup consisted of a few user programs that generated enough traffic to cause packet losses on the wireless interface. In fig. 4 we see a situation where bursty loss occurs. Since packets are dropped in bursts, a loss of around 0.2Mbps was noted every time a lossy situation occurs. The same setup is used with a class of applications that need distributed loss. We see that the packet drop is less (around 0.05Mbps) because the dropping of packets are spaced out in time. CBQ uses the tail drop mechanism wherein some packets are dropped in bursts while some individual packets are dropped too.

Further experimentation is under progress.

VII. CONCLUSIONS

In this paper we have described an architecture designed to support resource reservations in a mobile environment. The current implementation of the architecture used RSVP enhanced to signal passive reservations and CBQ enhanced to schedule passive reservations. The highlight of our approach is that we do not make assumptions about knowing in advance what path a mobile is going to follow. Also, we have studied QoS parameters specific to the mobile environment. An experimental testbed was developed to test the feasibility of the proposed approach.

The experiments show that resource reservation can be maintained as a mobile moves from one region to another. We also how that the "passive" reserved resources are not wasted if a mobile does not use the resources.

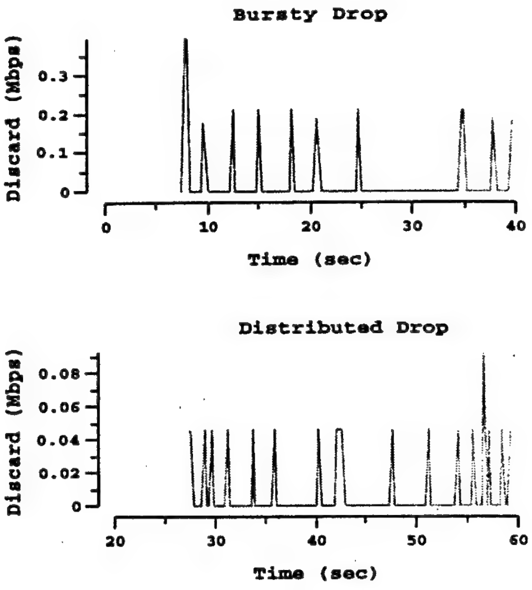


Figure 4: Bursty and Distributed loss based on loss profiles. When loss profiles indicates bursty losses, packets are dropped together (shown as loss of 0.2Mbps). When distributed loss is chosen, individual packets are dropped (shown by loss of 0.04Mbps). Note the difference in scale on y-axis.

ACKNOWLEDGMENTS

The authors would like to thank Dr. C. S. Raghavendra for his support.

REFERENCES

[1] L. Zhang, S. Deering, D. Estrin, S. Shenker and D. Zappala, "RSVP: A New Resource ReSerVation Protocol", *IEEE Network*, vol 7, pp. 8-18, Sept. 1993.
[2] R. Braden, L. Zhang, S. Berson, S. Herzog and S. Jamin, "Resource ReSerVation Protocol (RSVP) Version 1 Functional Specification", RFC 2205, Sept. 1997.
[3] A. K. Talukdar, B. R. Badrinath and A. Acharya, "MRSVP: A Reservation Protocol for an Integrated Services Packet Network with Mobile Hosts", Tech report TR-337, Rutgers university.
[4] S. Floyd and V. Jacobson, "Link-sharing and resource management models for packet networks", *IEEE Transactions on Networking*, Vol 3, No. 4, Aug. 1995, pp. 365-386.
[5] "WaveLAN Wireless Computing", www.wavelan.com.
[6] S. Singh, "Quality of Service guarantees in mobile computing", *Computer Communications*, Vol 19, No. 4, April 1996, pp. 359-371.
[7] "Bay Area Research Wireless Access Network", http.cs.berkeley.edu/~randy/Daedalus/BARWAN.
[8] "Wavelan driver on FreeBSD supporting roaming", www.monarch.cs.cmu.edu/wavelan.html.
[9] "RSVP code rel4.2a2" ftp://ftp.isi.edu/rsvp/release.
[10] "Code for alternate queueing. Version altq-0.4.2", www.csl.sony.co.jp/person/kjc/kjc/software.html.

In 30th IEEE SSST, March 1998

Next Generation ATM Communication Network in the Class Room

V.Rajaravivarma
Electronics & Computer Technology
North Carolina A&T State University
Greensboro, NC 27411
Email: veeramu@ncat.edu

Krishna Sivalingam
Electrical Engineering & Computer Science
Washington State University
Pullman, WA 99164
Email: krishna@eecs.wsu.edu

Abstract –

This paper describes the education and improved curricular content of computer networking and communications courses through the introduction of hands-on ATM and wireless network programming to students. The School of EECS at Washington State University (WSU) and School of Technology of N. C. A&T University (A&T) are in the process of establishing two laboratories to teach advanced computer networking. Traditional computer networking courses primarily tend to provide students with hands-on software development and network performance experience with TCP/IP networks. These labs plan to enhance the expertise of students through introducing next-generation networks. In particular, projects and experiments based on ATM networks and wireless networks will be introduced, both of which are increasingly important technologies in the industries students hope to join. The objective is to prepare the students for both industry and advanced graduate research. Two different approaches will be used to teach practical advanced networking concepts. At WSU, a project-based approach will be used where the students focus on writing software applications using the ATM and wireless application programming interface (API). A&T will follow an experiment-based approach where each course will typically involve two hours of theory and two hours laboratory work.

1 Introduction

The purpose of this paper is to describe the introduction of advanced networking concepts including wireless and Asynchronous Transfer Mode (ATM) networks in undergraduate courses. The courses will be taught at Washington State University (WSU) and at North Carolina A&T University (A&T). Two separate wireless and ATM network laboratories are being established using funding from the National Science Foundation through an Instrumentation and Laboratory Improvement (ILI) grant. Undergraduate and graduate students will be able to obtain hands-on learning experience with wireless and ATM networks and compare that to their experience using current TCP/IP and Ethernet networks.

Each laboratory will consist of a set of computers that are interconnected by a Fore ATM switch running at 155 Mbps. In addition, the WSU lab will contain a wireless local area network that is based on Lucent Technologies WaveLAN products running at 2 Mbps. The computers will also be connected to the departmental Ethernet network and to the rest of the Internet.

The rest of the paper is organized as follows. Section 2 explains the need and the motivation for establishing the wireless and ATM network teach-

[1] This work was supported by the National Science Foundation under Grant DUE-9751198 and the United States Air Force Office of Scientific Research under grant F-49620-97-1-0471.

ing laboratories. Section 3 presents the organization of the laboratories and describes the equipment. Section 4 describes the courses and the teaching methodology based on projects and experiments. Section 5 describes the current status of the project. Section 6 summarizes the paper.

2 Motivation

The motivation for establishing these advanced laboratories is described in this section.

In a typical undergraduate course in networking, the fundamentals of networking and introduction to programming in TCP/IP are covered [10, 9, 3]. Topics including Open Standards Interconnection (OSI) hierarchical protocol architectures and application programming using BSD Sockets [2] or WinSock interface are usually taught. Since TCP/IP networks are available in most departments, teaching TCP/IP programming is quite common. However, there is a need to teach advanced networking concepts such as ATM and wireless networks to undergraduates. These networks are moving from research labs and test facilities to become more mainstream in the industries that students will join.

Asynchronous Transfer Mode (ATM) networking was chosen as the means to deliver Broadband Integrated Services Digital Networks (B-ISDN) [8]. Its main goals are to support multimedia traffic, deliver quality-of-service, and offer faster cell-based switching. The current status of ATM network deployment is such that ATM looks like the most likely candidate for back-bone switching. However, ATM does not appear poised to capture the desktop market where Ethernet, Switched Ethernet, and Gigabit Ethernet may ultimately succeed. However, it is essential that students understand and gain experience in both ATM-based networks and current TCP/IP-based networks.

An introduction to the basics of ATM is now found in recent networking textbooks such as [10]. What is lacking is providing the students with actual hands-on experience in native ATM Application Programming Interface (API), IP over ATM, and in learning the differences between ATM and TCP/IP based networks. The knowledge gained from application development will enable them to better understand the underlying networks and protocols. It will also enable them to better understand existing networks based on Ethernet and TCP/IP, and appreciate the design differences between the different networks. This motivation has led to the establishment of a teaching laboratory that contains a set of computers interconnected through an ATM switch.

Another type of network that is becoming more commonplace is the wireless/mobile network. Wireless communication has been through two generations of development [7]. The first generation of wireless networks supported typically voice only communication. The second generation of wireless networks support both voice and data using networks such as CDPD, ARDIS, WaveLAN, and so on [7]. The third generation promises support for multimedia with quality-of-service provided, and easy WWW access.

Wireless communications courses typically tend to focus more on the wireless channel and communication characteristics. There is a growing body of research that is addressing the impact of wireless channel on higher-level protocols [1]. Students should gain experience with developing applications and understanding network performance in wireless networks to better understand them from a networking protocol perspective. This was the motivation to include a wireless local area network in the laboratory.

3 Laboratory Setup

This section describes the equipment and organization of the two laboratories. A schematic of the network laboratory at WSU is shown in Fig. 1.

The lab consists of at least six computers connected to a Fore ATM switch. The ATM switch has 12 155 Mbps ports and one 622 Mbps port. The 622 Mbps port enables experiments and projects that can explain the issues involved in very high-speed networking. The machines that will be connected to the ATM switch are a SPARC Ultra server, and six PC clients running Linux/Windows NT. The wireless network is set up using Lucent Technologies' WaveLAN. The wireless net-

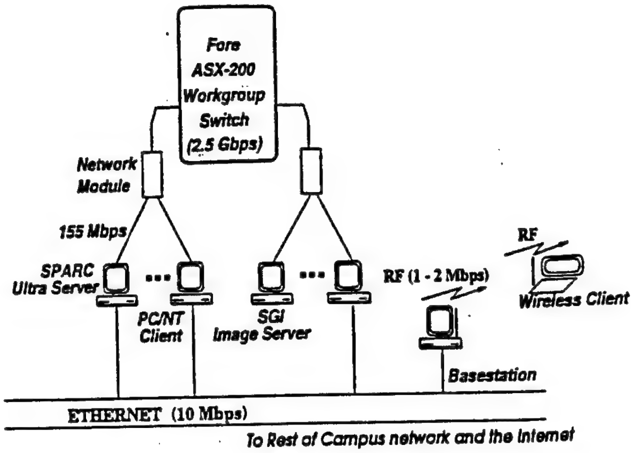


Figure 1: Schematic of the ATM and wireless network laboratory.

work basestation is accomplished by attaching the basestation wireless equipment to a PC running Linux. This basestation will also be connected to the departmental Ethernet network as shown in the figure. The wireless clients will be laptop computers fitted with the radio interface. The clients will communicate with the rest of the network through the basestation.

The laboratory at A&T will consist of at least twelve personal computers connected over an ATM network. One of the computers will be the image server and will contain the network management software. The experiments taught here will focus on image processing applications using the high-bandwidth network. A proposal to add wireless equipment to study image transmission over wireless links is under consideration here.

4 Teaching Methodology

The following describes the courses that will use and the teaching methodology at the two institutions. WSU will teach using a set of projects each expected to take 2-4 weeks long. A&T will develop a set of experiments that can be completed within a week of laboratory time. The techniques will derive from strategies that have been adopted in other networking courses [4].

The courses currently offered at WSU that will make use of the requested equipment include:

(i) CS/EE 455: Computer Networks. This is an undergraduate course in networking which will introduce students to TCP/IP, ATM and wireless networking. Course projects will be designed to provide them the fundamentals of networking in these advanced technologies.

(ii) CS/EE 555: Computer Communication Networks. This is a graduate course which is also available to undergraduates. The students are taught analytic techniques to understand the design decisions and performance of computer networks. A term-long project, done in teams, is required as part of the course. The students are given a wide variety of relatively open-ended projects to choose from. The idea is that these projects might develop into M.S. projects and theses.

CS 455 will use the ATM switch and require the students to write network applications using the TCP/IP Socket interface and Fore ATM API interface. Traditionally, undergraduate networking courses are taught from the lower layers (physical) upwards to the transport and application layers. A different technique is used here. The OSI architecture is taught first, followed by a discussion of the application and transport layer protocols and interfaces. This enables the students to immediately start writing application programs and get comfortable with the concepts. This method was adopted at a course the author taught while at University of North Carolina and has observed good success. During this time, the students were asked to write a subset of ftp client and server which they started very early in the semester.

At WSU, three to four course projects are assigned each semester for the undergraduate course and an open team project for the graduate course. The student class and term projects are designed to enable the students to better understand ATM and wireless networking concepts. The experiments will also be designed to drive home the issues arising from integrating networks each operating at a different speed. The current departmental Ethernet network runs at 10 Mbps, the ATM network will operate at 155 Mbps, and the wireless network will operate at 2 Mbps. This diversity in network speed and technology will teach the students how different applications and protocols have to be designed for high-speed networks and for lower speed wireless networks. The projects are chosen with the following key goals in mind:

1. Teach students the basics of connection-less and connection-oriented networking. TCP/IP and Wireless provide the connection-less network while ATM provides the connection-oriented network [10].

2. Investigate the reliability problems introduced by the wireless environment and design applications that adapt to such a network. Adaptive applications that change to network conditions will be the key to future mobile applications [5].

3. Introduce the importance of quality-of-service and resource allocation in networks. For example, a voice or video application will be designed with TCP/IP and wireless with no bandwidth guarantees, and with ATM with bandwidth reservation.

4. Investigate the impact of wireless link characteristics on TCP/IP protocol stack [1].

5. Teach students the differences between a broadcast environment such as Ethernet versus a switched environment such as ATM.

6. Introduce how legacy applications such as those based on Ethernet and IP will have to be supported over new technology such as ATM using techniques such as LAN Emulation and IP over ATM.

7. Teach students how to compare the performance of applications based on TCP/IP, ATM and Wireless.

8. Teach students the basic principles of network management using Fore Systems' ForeView network management software. Managing the network is a critical operation and it is essential that the students are trained to efficiently manage networks in addition to writing network application software.

The students at A&T will be prepared to understand the basic principles of data communication. Network management, system installation, and maintenance of the hardware and software will be the focus of the laboratory experiments. The experiments will be designed to familiarize the students with image data transfer and applications.

Medical imaging applications require transfer of large amounts of data and high-speed ATM networks appear to be a possible solution. There are three major hospitals and medical research institutions in the vicinity of this university, which are currently involved in medical imaging applications. Studies have shown that physicians are willing to wait about 1.6 seconds between images, and a ATM network at Duke University has accomplished image transfer in 1.5 seconds [6]. This is an example of applications that will be more widespread in the near future. The goal of the laboratory experiments is to train the students with such high technology.

The courses taught at NC A&T university that will use this lab include:

(i) ECT 620 Telecommunications management: This course teaches fundamental principles of telecommunications management, which includes network management and administration, the telecommunication marketplace and the planning and evaluation of systems.

(ii) ECT 630 Electronic Communication networks: This course involves an intensive study of the principles involved in designing Local, Metropolitan, and Wide Area Networks. The student will be required to design an appropriate network to meet predetermined specifications.

The laboratory also provides the platform for more applications such as distributed image processing. The faculty member at A&T involved in this project has published work on image decomposition and processing problems. The idea here will be to decompose an image into sub-images and distribute the sub-images to different machines connected over the ATM network.

The advantages of setting up two separate laboratories at the two universities are: (i) WSU students will be taught using project-based techniques, and A&T students will be taught using experiment-based techniques. The students can derive greater benefit through our exchange of experience from both pedagogical techniques; and (ii) the students will also investigate internetworking through data transfer from an ATM switch on one campus to the other remote ATM switch over the Internet. A long range goal is to accomplish video transmission between the two sites and provide live on-line lectures.

5 Current status

At present, both universities are engaged in procuring the equipment for establishing their respective laboratories. At WSU, CS 455 will be offered in the Spring 1998 semester and CS 555 in the Fall 1998 semester. The results from the student completion of the projects will be evaluated using feedback obtained from the students. The feedback will be based both on directed and open questions. A comprehensive evaluation of the relative merits of the project and experiment based teaching methods will be conducted. The projects and experiments that were developed will be available as course materials over the World Wide Web at our university sites and also through NSF's computer science course repository.

The results from the courses and the evaluation of the project will be published in subsequent papers. At this time, the goal of this paper is to disseminate the nature of the laboratories and the proposed teaching methods.

6 Summary

This paper describes an attempt to teach advanced networking concepts such as wireless and ATM networking in undergraduate courses. Two laboratories containing ATM and wireless equipment are being established at Washington State University and North Carolina A&T University. Two different teaching techniques will be adopted to provide hands-on experience: project based and experiment based. Students will be taught the different networking concepts and understand the fundamentals of ATM, wireless and current TCP/IP networks.

References

[1] BALAKRISHNAN, H., PADMANABHAN, V., SESHAN, S., AND KATZ, R. H. A comparison of mechanisms for improving TCP performance over wireless links. *IEEE/ACM Transactions on Networking* (June 1997).

[2] COMER, D. *Internetworking with TCP/IP, Vol III: Client-Server Programming and Applications, Socket Edition*, 2 ed. Prentice-Hall, 1996.

[3] COMER, D. *Computer networks and Internets.* Prentice-Hall, 1997.

[4] FINKEL, D., AND CHANDRA, S. Netcp - A Project Environment for an Undergraduate Computer Networks Course. In *Technical Symposium on Computer Science Education* (Mar. 1994).

[5] HOKIMOTO, A., AND NAKAJIMA, T. Handling Continuous Media in Mobile Computing Environment. In *Proc. IEEE Intl. Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV '96)* (Apr. 1996). pp. 175–182.

[6] MAYER, J. H. ATM networks transfer medical images at high speeds. *Vision Systems Design* (Oct. 1996). 28–31.

[7] PAHLAVAN, K., AND LEVESQUE, A. H. Wireless data communications. *Proceedings of the IEEE 82*, 9 (Sept. 1994), 1398–1430.

[8] PRYCKER, M. D. *Asynchronous Transfer Mode: Solution for Broadband ISDN*, 3 ed. Prentice Hall, 1995.

[9] STALLINGS, W. *Data and computer communications*, 3 ed. Macmillan, 1991.

[10] TANENBAUM, A. S. *Computer Networks*, 3 ed. Prentice Hall, Mar. 1996.

IN IEEE ICPWC, DEC. 1997

MAC Layer Scheduling Strategies during Handoff for Wireless Mobile Multimedia Networks

Shalinee Kishore[1], Prathima Agrawal[2], Krishna M. Sivalingam*,[3] and Jyh-Cheng Chen[4]

[1]WINLAB, Rutgers University, Piscataway, NJ 08855

[2]Networked Computing Technology Department, AT&T Labs, Whippany, NJ 07981

[3]School of Electrical Engineering & Computer Science, Washington State University, Pullman, WA 99164

[4]Department of Electrical & Computer Engineering, State University of New York at Buffalo, Buffalo, NY 14260

Abstract:

Mobile multimedia systems must - amongst other things - account for user mobility in its network architecture. In this paper, we present strategies for accommodating continuous service to mobile users via modifications in the design of the multiple access (MAC) protocol. More specifically, we focus here on the adjustments required in the scheduling mechanism operating at basestations which grants access to mobile terminals by assigning them channels and time slots. In order to meet the needs to such mobile users, we propose here a scheduler that uses anticipatory handoff indications from neighboring basestations to pre-assign slots for mobiles that are close to entering its coverage area. The paper examines the considerations necessary in pin-pointing such potential handoff mobiles. Additionally, an alternate MAC frame structure is presented to accommodate these potential users. Parameters that improve the reliability and utilization of these assignments are also studied.

1 Introduction

The two recent trends in telecommunications, i.e. wireless personal communications and broadband networking for multimedia information services, aim to come together in the next generation of wireless networks [1,2]. Forecasts of multimedia integration into wireless networks has driven recent research to develop wireless architectures to accommodate *seamless* or uninterrupted service while ensuring Quality-of-Service (QoS) transmission of multimedia traffic. Medium access control (MAC) protocols designed for such wireless multimedia networks employ efficient scheduling algorithms as a means to maintain QoS guarantees to individual network sessions [3]. Wireless ATM is an example of such a network [2]. In mobile multimedia networks, the handoff of users from one serving basestation (BS) to another imply additional tasks for BS schedulers. In addition to computing fair and optimal schedules for current users that meet each terminal's QoS requirements, an efficient scheduling algorithm designed for such a mobile network must also quickly and effectively update schedules in appropriate neighboring cells so as to provide QoS guarantees during and after handoff.

Schedulers for such mobile multimedia systems must then use information on resource demand of current users and potential entering users to compute the schedule for their cells. A mobile terminal appears to each BS as a traffic source with a particular request characterization – determined by traffic-type (voice, video, or data), QoS requirements (such as packet delay, jitter, or throughput), and buffer status. The MAC layer must then process this request and schedule transmission slots during the MAC frame for each mobile both currently in and potentially entering its cell region. Furthermore, to conserve resources a departing mobile terminal's scheduled slot in the previous serving BS must be promptly cleared and made available for other users in that cell. Thus scheduling algorithms and therefore MAC protocol designs for mobile multimedia networks must incorporate mechanisms to account for user mobility.

A reservation and scheduling based access protocol such as described in [4] is considered here. Transmission is organized into frames, where each frame consists of at least a reservation phase and an uplink (mobile-to-BS) data phase. In the reservation phase, mobiles place requests for transmission or update of the queue status of the individual connections. The BS then executes a scheduling algorithm to allocate slots to the mobile's connections during the data phase.

In this paper we present possible strategies to incorporate handoff considerations in the MAC layer, specifically in the scheduling algorithms for mobile multimedia networks. The mechanism that we present here uses notification from the higher levels and lower physical level to develop an alternate MAC frame structure to deal with terminals in or *close to* handoff. This new MAC structure now includes a separate ***handoff phase*** during which these mobiles from neighboring cells are allocated slots for transmission.

The proposed mechanism works as follows. A region is established at the outskirts of the cell region that serves as the basis for determining which mobiles are ***close to*** handoff. In this proposed system, the BS uses measured received signal strength from a transmitting mobile to approximate if the terminal is in this boundary region. Once determined

*This work was supported in part by Air Force Office of Scientific Research grant F-49620-97-1-0471. Part of the research was conducted while this author was at AT&T Labs and University of North Carolina at Greensboro

to be in the region, the terminal then becomes part of an anticipated handoff procedure. Using periodic updates, the current basestation (BS) notifies its neighboring basestations of those terminals that lie in this boundary region.

The new BS begins negotiation of QoS at the transport layer in anticipation of an impending handoff of an incoming mobile. Once notification from the higher levels is received, the new basestation(s) places the terminal in its schedule, more specifically in the new, dynamic *handoff phase* of the MAC frame. Similarly, the current BS also marks the mobile as a possible departing terminal and places its scheduled slots in its handoff phase as well. As the mobile moves closer to the cell boundary and a hand-over of the terminal is finally achieved, the current BS immediately clears the terminal's slots from its handoff phase and the new BS simultaneously incorporates the terminal's traffic into its schedule without any waiting time. In anticipation of a handoff the BS could possibly schedule transmission times for users who may not enter its coverage area. The paper also focuses on means to reduce these idle or wasted assignments.

The paper is organized as follows. Section 2 summarizes previous work on MAC protocols and handoffs that motivated the discussions here. Section 3 describes the overall system architecture and the specifics of the MAC protocol under consideration. In Section 4 we present our proposed strategies to extend handoffs into the MAC level scheduling mechanism. Section 5 describes the future direction of this research.

2 Previous Work

A great deal of the discussion on allocating bandwidth to mobile users in wireless networks has thus far focused on higher level protocols with little emphasis on their effect on the MAC layer. In [5], Singh proposed to extend the definition of QoS in mobile systems to include a guarantee of seamless service as well as an assurance of graceful degradation of service in situations where resource demands exceed network capacity. Oliveira, Kim, and Suda [6] presented a QoS guarantee algorithm for high-speed multimedia wireless networks in which bandwidth reservation was used to achieve uninterrupted communications. In [7], efficient channel management schemes were proposed for cellular networks in the presence of handoffs. More recently, Das, Jayaram, and Sen [8] have presented an approach called *bandwidth compaction* as a more efficient means to provide bandwidth resources to mobile users. Another higher-layer implementation designed to accommodate handoffs is a *virtual connection tree* proposed by Acampora and Naghshineh [9]. This is primarily a scheme for setting up connections or routes through the backbone network where routes change frequently due to handoffs. The central objective of all these techniques is to design protocols that can overcome the changing network conditions from one cell to the next and provide a mobile with seamless communications.

All these schemes, however, approach handoff from the higher layers of the protocol stack. In such proposals, seamless service is guaranteed to a user by reserving network resources, such as bandwidth or virtual circuit connections. After all the QoS tweaking and bandwidth shuffling, the mobile appears at the MAC layer as a traffic source with a particular traffic-type, a buffered queue, and QoS requirements - such as the packet delay, jitter, throughput, etc. As indicated in [3], literature on such wireless networks as RATM, WATMnet, SWAN, etc. examine the MAC layer in terms of the different multiple access techniques or in terms of dynamic protocols that allow for variable bit rate transmission with QoS considerations. Additionally, the analysis of these protocols is implemented using simplified scheduling algorithms that take into consideration QoS guarantees, traffic characteristics, and channel conditions. The simulation results presented in this paper indicate that such scheduling schemes based on round-robin, first-come-first-serve (FCFS), or priority principles do not provide fair and efficient transmission of multimedia traffic with QoS guarantees. The service time performance suffers because of the increased number of retransmissions in wireless communications, and the limited bandwidth of such networks in turn limits the ability of the scheduler to deal with the priority of different multimedia traffic types. This work points to the need for appropriate scheduling algorithms that can better accommodate the limited bandwidth and error-prone channel inherent in wireless links.

What we propose in this paper is an additional dimension to these scheduling concerns for wireless networks. Not only do the slot-by-slot assignments require a more robust scheduling mechanism to deal with QoS guaranteed transmissions, but the frame-by-frame scheduling of the MAC frame must also incorporate another characteristic inherent in such wireless networks, *user mobility*. A crucial shortcoming of simple MAC-level schedulers proposed for mobile networks lies in their inability to accommodate handoffs. As pointed out in [3], robust scheduling at the MAC level seems to be the most appropriate approach if such QoS requirements are to be met. Therefore to provide reliable performance to mobiles in transit while maintaining the higher-layer requirements of a guarantee of seamless service, BS schedulers must account for both present and potential users when computing the frame-by-frame schedules for their cell region.

3 System Descriptions

The paper considers an infrastructure-based wireless network depicted in Figure 1. A geographic coverage area for the wireless network is composed of a collection of smaller regions called *cells*, represented by the hexagons. A cell can be different sizes - macro, micro, or pico - depending on the network. Each cell is serviced by a which in turn serves a set of mobile terminals. A *Mobile Switching Center* coordinates the activities of a number of neighboring basestations and connects these basestations and their mobile terminals to fixed networks like PSTN, ISDN, or an ATM backbone.

A mobile can originate and terminate multiple data connections, that enable it to communicate with other computers and communication devices. All communication to and from the mobile is through the BS. Each such connection

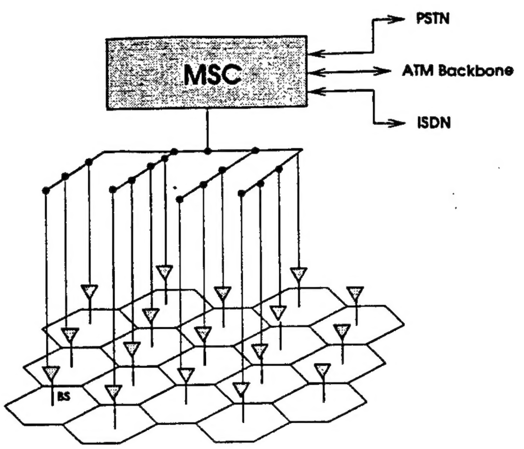


Figure 1: Infrastructure for wireless network

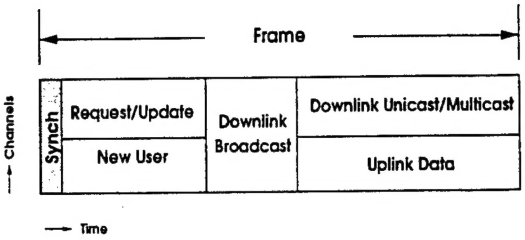


Figure 2: General MAC frame structure

is referred to as a Virtual Circuit (VC). This technique is adopted in ATM (Asynchronous Transfer Mode) networking for multimedia communication [10]. Each VC is associated with a transmission priority established by the mobile application utilizing this VC for communication. These priorities are utilized by the BS when allocating channels to the mobiles. Each mobile maintains a separate queue for each of its VC. Information arrives at each queue in the form of a *packet* and is buffered until transmission.

Transmission in the network at the MAC layer is organized into *frames* which is further divided into subframes which in turn are comprised of slots. Each mobile uses a channel - either a carrier frequency as in FDMA or a CDMA code - to communicate with the BS during a particular slot. Figure 2 shows a frame with three subframes. The following is a description of the frame structure.

1. At the beginning of each frame, there is a frame synchronization phase that aids new and current users to establish and maintain synchronization.

2. In the *request/update and new user*, mobiles transmit their current transmission requests such as their queue status to the BS.

   Note that separate sets of channels are allocated to new users and for current users during this phase. The new users learn of the appropriate codes to use during the synchronization phase.

3. Next comes the *downlink broadcast* phase.. Downlink refers to communications directed from the BS to the mobile. In this phase the BS broadcasts data, acknowledgments, and scheduling information that all mobiles need to receive. Using this schedule, the mobiles know exactly during which slots they can transmit to the BS.

4. The *downlink unicast/multicast and uplink data* phase follows. During this data phase a group of downlink channels are used so the BS may transmit unicast or multicast data on different channels. During this phase mobiles can also transmit data on the uplink.

The proposals presented in this paper all stem from modifications to this MAC frame structure, more specifically to the fourth item listed above, the data phase. The cell architecture described here is used to pinpoint those mobiles that may cross from one cell to the next. The specifics of these proposals are presented in the next section.

4 Proposed Strategies

This section describes our proposed strategies in terms of the departing region and the MAC frame structure. The departing region helps identify the users that may potentially enter the coverage area of a particular BS, and the MAC frame structure is adjusted so that it may account for these potential terminals.

4.1 Departing Region

We first provide a strategy to determine how a mobile currently registered at one BS is classified as a potential new user for a neighboring BS. Here we adopt the approach presented in [8]. A *departing region* is defined on the outskirts of the cell, as shown by the shaded region in Figure 3. The BS then determines which of its current mobiles lies in the departing region. A detection process based on the user terminal's received signal strength (RSS) can be used to determine with a certain degree of probability if a particular mobile lies in this peripheral region. Once this is done and a mobile is marked as *departing*, the BS can also determine towards which two neighboring cells the mobile is headed. This is established either via the use of sectorized antennas or by taking RSS measurements at the six neighboring basestations. Using this process under the supervision of the MSC, the marked basestations can then begin modifying their schedules to accommodate this mobile terminal.

The size of this departing region is an important parameter in this implementation. This size is effectively interpreted from the distance, D, indicated in Figure 3. If D is small, then more mobiles will be marked as *departing*. Chances are, however, that not all these mobiles will leave the current cell region in which case the neighboring cells will unnecessarily adjust their schedulers. This can in turn lead to wasted resources at these basestations which is highly undesirable during heavy traffic load. If, on the other hand, the distance

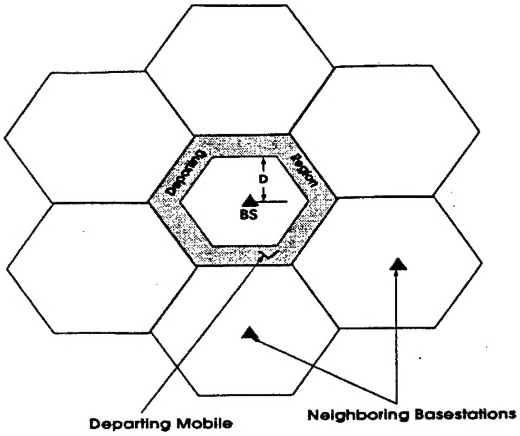


Figure 3: Departing Region to determine departing mobiles and appropriate neighboring cells

D is too large, then there might not be enough time between the instance the mobile enters the departing region to when it crosses into the neighboring cell for the basestations and the coordinating MSC to (1) realize the mobile as *departing*, (2) determine and inform the neighboring basestations, and(3) allocate the necessary scheduled slots in the ***uplink data*** phase to the mobile. Such a scenario stands a greater chance of occurring when the mobiles in the prescribed wireless network travel at high speeds. The macro and often the micro cellular architectures encounter such high mobility speeds. Thus the parameters required to compute the distance D are:

1. The traffic load at the current and neighboring basestations.

2. The scheduling-set-up delay, T, which is:

$$T = T_d + T_n + T_s \qquad (1)$$

   where T_d is the time needed to detect if a mobile is *departing*. T_n is the time needed to determine and inform the appropriate neighboring cells, and T_s is the delay in scheduling the mobile in the neighboring cells.

3. The size of the cell region.

4. The average speed of the mobiles in the network.

5. The propagation characteristics of the cell region to determine the RSS values that appropriately form the departing region.

4.2 MAC Frame Structure

Next, we focus on how the MAC frame structure has to be adjusted, or more specifically how the scheduler has to modify the frame structure to accommodate potential new users. Here we propose that the data phase of the MAC frame described earlier must now be adjusted to appear as in Figure 4.

During the data phase, the BS uses downlink channels to unicast or multicast information to the terminals. Simultaneously, the mobile terminals employ the uplink channels and the schedule received during the ***downlink broadcast*** phase to transmit their buffered packets during their assigned slots. For both the downlink and uplink transmissions, the scheduler must now assign a certain number of its slots in the data phase to those mobiles that have been marked as *departing* in a neighboring cell and are potentially headed towards the current region. The transmission requests and other VC parameters of these mobiles are conveyed to the current BS scheduler via the MSC in a periodic fashion. Thus, the MSC informs the current BS of the transmission requests of all mobiles potentially headed towards its cell region. With knowledge of the traffic types, QoS requirements, and buffered queues of these mobiles, the scheduler then incorporates these requests into its algorithm and assigns the potential terminals slots during the data phase. Thus as shown in Figure 4, the data phase for both the downlink and uplink channels is divided into two sub-phases: one for current users and one for potential handed-off incoming mobiles.

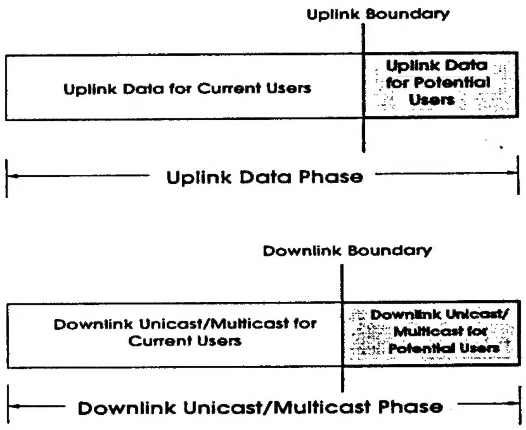

Uplink Boundary
Uplink Data for Current Users
Uplink Data for Potential Users
Uplink Data Phase
Downlink Boundary
Downlink Unicast/Multicast for Current Users
Downlink Unicast/ Multicast for Potential Users
Downlink Unicast/Multicast Phase

Figure 4: Data Phase that incorporates potential users

The boundary between the two subphases is dynamically adjusted after the MSC updates the current BS concerning the transmission requests of the *departing* mobiles in the neighboring cells. The number of slots allocated to these new potential users depends on their traffic demands. However, simply reserving slots that meet the requirements of the VC's of each potential hand-off user implies wasted slots in perhaps many MAC frames - especially if the MSC updates occur in large intervals. Slots can be assigned to these potential users and could go idle, i.e. wasted, if the user does not move into the new cell region before the next MSC update or terminates the call before crossing the cell edges.

A more efficient reservation scheme must therefore also consider the following:

1. *Higher layer mobility management and location estimation information.* This information can then be used to realize the probability of a particular mobile to actually become a hand-off user in the current cell. If, for example, location estimation is able to place a mobile on a free-way headed towards a region which is not the current cell's, then that user's transmission requests could be deprioritized, if not removed from the scheduler's consideration altogether. Similarly, if these upper layer processes confirm the path of the mobile is towards the current cell, then the scheduler can assign slots with a great probability of utilization before the next MSC update.

2. In order to uphold the performance of registered mobiles in the cell, the reservation scheme also has to depend on the *traffic load of the VC's of current terminals.* The scheduler cannot simply allocate slots to users who may or may not appear in their region during those intervals when the current users have substantial transmission requests. For example, during heavy traffic load, the scheduler should accommodate the potential users with a coarser granularity, i.e. by incorporating a hand-off subphase only once in ever few frames instead of every frame.

3. The scheduler must also use information from *past history* to allocate slots in the present MAC frame. The scheduler must first consider the current BS's blocking statistics during a past interval before reserving slots for potential users. If, due to the hand-off subphase, the BS has started to block too many new terminals in its region, then the granularity of this subphase must be adjusted as described above. Furthermore, the scheduler must also examine - within a past interval - the utilization of slots in the hand-off subphase. In comparing this utilization to the transmission requests of the potential users in the past, the scheduler can better estimate the actual required number of slots in the present hand-off subphase.

4. Finally, a more efficient scheduler must have *means to reclaim idle slots.* Thus mechanisms must be incorporated to further minimize - if not completely eliminate - wasted slots during this proposed hand-off subphase.

5 Further Work

We are currently developing a set of strategies to identify the percentage of slots (denote this β) that is reserved for the *handoff* phase. This parameter has a significant impact on the performance of the system. Note that there is a potential tradeoff here between lower dropping probabilities for incoming handoff connections versus lower blocking probabilities for new connections originating in the current cell. A higher value of β favors incoming handoffs and results in lower dropping probabilities for these connections. A lower value of β, favors new connections originating in a cell. This is currently under investigation using simulation and analytic methods. Once a set of such strategies that provide near-optimal or optimal performance have been identified, system performance with realistic traffic mobility patterns will be studied.

Acknowledgments

The authors wish to thank Prof. Paramesh Ramanathan of the Department of ECE, University of Wisconsin, Madison for valuable discussions that significantly improved the quality of the paper.

References

[1] M. Naghshineh (Guest Ed.). Special issue on wireless atm. *IEEE Personal Communications*, 3(4), August 1996.

[2] T. R. Hsing, D. C. Cox, L. F. Chang, and T. Van Landegem (Guest Ed.). Special issue on Wireless ATM. *IEEE Journal on Selected Areas in Communications*, 15(1), January 1997.

[3] Jyh-Cheng Chen, Krishna M. Sivalingam, and Raj Acharya. Comparative analysis of wireless ATM channel access protocols supporting multimedia traffic. *ACM/Baltzer Mobile Networks and Applications*. To appear.

[4] K. M. Sivalingam, M. B. Srivastava, P. Agrawal, and Jyh-Cheng Chen. Low-power access protocols based on scheduling for wireless and mobile ATM networks. In *Proc. IEEE International Conference on Universal Personal Communications (ICUPC)*, San Diego, CA, USA, October 1997.

[5] Suresh Singh. Quality of service guarantees in mobile computing. *Computer Communications*, 1(19):359–371, April 1996.

[6] C. Oliveira, J.B. Kim, and T. Suda. Quality-of-service guarantee in high-speed multimedia wireless networks. In *IEEE International Communications Conference*, pages 728–734, Dallas, TX, USA, 1996.

[7] P. Agrawal, D. Anvekar, and B. Narendran. Channel management policies for handovers in cellular networks. *Bell Labs Technical Journal*, 1(2):97–110, Autumn 1996.

[8] S. Das, R. Jayaram, and S. Sen. An optomistic quality-of-service provisioning scheme for cellular networks. In *DCS*, 1997.

[9] A.S. Acampora and M. Naghshineh. Control and quality-of-service provisioning in high-speed microcellular networks. *IEEE Personal Communications Magazine*, 1(2), Second Quarter 1994.

[10] M.D. Prycker. *Asynchronous Transfer Mode: Solution for Broadband ISDN.* Prentice Hall, 1995.